Project ref. no.	*IST-1999 11876*
Project title	**CORETEX - Improving Core Speech Recognition Technology**

Deliverable status	*Public*
Contractual date of delivery	t0+18
Actual date of delivery	t0+24
Deliverable number	*D2.2*
Deliverable title	*Report on Acoustic Meta-data Mark-Up*
Type	*RE-Report*
Status and version	*First version*
Number of pages	*48*
WP contributing to the deliverable	*WP2*
WP / Task responsible	*CUED*
Author(s)	*P. Woodland, M. Gales, J. T. Karunatillake, D. Yuan (CUED); M. Federico, M.Cettolo (IRST); J.L. Gauvain, L. Lamel, G. Adda, F. Lefevre, H. Schwenk (LIMSI); H. Ney, M. Pitz (RWTH)*
EC Project Officer	*Domenico Perrotta*
Keywords	*core speech recognition, acoustic meta data, segmentation, speaker tracking, acoustic environment, speech clustering, confidence levels*
Abstract (for dissemination)	This report overviews the research carried out during the first 18 months of the CORETEX Project concerning Acoustic Meta-Data markup. The research concerns the automatic generation of information other than a word level transcription that can be automatically extracted from the acoustic data. This enriched information from the audio stream can be used to improve recognition and in other useful contexts. Techniques and results are presented for the markup of sections of speech, non-speech music, non-speech noise, speaker, background environments, speech in varied acoustic environments, and confidence rates in recognition at word level.

Contents

1 Introduction

The majority of state-of-the-art speech recognition systems simply generate a word level transcription as the output. This lacks many levels of information that would be useful to someone, or some automatic process, reading the transcription. These additional levels of information, commonly referred to as *meta-data* may be split into two broad classes.

- **Acoustic meta-data**: This consists of information that may be directly extracted from the acoustic data. Examples of this are acoustic environment, channel condition, speaker information such as gender and possibly speaker identity if enrolment data is available.

- **Linguistic meta-data**: This consists of making use of higher level information. Examples consists of capitalization and punctuation, topic boundaries and summarisation.

This report describes the initial work performed under the CoreTex project for acoustic mark-up. The report addresses various issues associated with the generation of meta data, supported by experiments on Broadcast News data.

- **Acoustic segmentation**: For tasks such as broadcast news transcription the data arrives in the form of a continuous single stream of acoustic data. The first stage of processing is to clean the data stream by identifying speech and non-speech sections, and discarding the non-speech sections. The second stage of processing is to segment the cleaned stream into homogenous blocks of data. The correct number of segments to partition the data into is unknown a-priori. Section 2 describes experiments on broadcast news transcription data to address these problems of segmentation.

- **Speaker tracking and channel labelling**: Having segmented the data into (hopefully) homogenous blocks, the segments must then be labelled with the appropriate speaker in the foreground, and acoustic environment in the background. These labelling tasks are discussed in section 3.

- **Cluster assessment**: After partitioning the single stream of data into a series of segments, the segments are clustered. It is important to determine how well a clustering system has performed and what kind of clustering has been achieved. For this clustering schemes must be evaluated with respect to multiple possible sets of labelling schemes, for example labelling schemes that consider speakers in multiple acoustic environments. Section 4 describes a modified scoring criterion which aims to address this problem. The techniques are applied for speaker-environment tracking for labelled homogeneous segments, and speaker turn assignment for unlabelled automatic segments.

- **Confidence levels**: An additional form of acoustic meta-data is the confidence level of individual words being correct. Details of this work is given in 5.

2 Acoustic Segmentation

When applying automatic speech recognition to broadcast news data, a segmentation step must be performed in advance. The goal of this segmentation stage is, in the first instance, to partition the whole audio stream into speech and non-speech segments and, in the second, to cut the speech segments into reasonably short frames while discarding the non-speech portion.

This section details two techniques appropriate for the segmentation tasks. The success of using a Maximum Mutual Information decoder for the pre-segmentation task of separating speech from non-speech sections in German Broadcast News is demonstrated in section 2.1. The process produces acoustic markup as speech, non-speech music, or non-speech noise, and also produces clean speech segments of high purity.

Secondly, several model selection criteria, which perform transition detection, are evaluated for the task of segmenting cleaned speech in section 2.2. The aim is to detect spectral changes that occur within the signal that are mainly due to channel and source switches. The signal is segmented at these points of spectral change. It is demonstrated that several of the techniques can be fine tuned to achieve high precision segmentation of Italian Broadcast News.

2.1 Discriminative Audio Stream Segmentation and Labelling on German Broadcast News

For this task audio stream segmentation is considered a specific decoding task with only a small number of classes. In this instance, the respective model complexities are typically very low and a segmenter may benefit from discriminative training methods.

Discriminative training criteria such as Maximum Mutual Information (MMI) and Minimum Classification Error (MCE), often outperform maximum likelihood (ML) based training methods, especially for low model complexities. So far, discriminative training techniques were used to optimize speech recognition systems at the model level. Nevertheless, due to their general formulation they can also be used for segmentation tasks. In order to mark-up audio streams, the MMI criterion was applied to train a segmenter using a generalized probabilistic gradient descent method for parameter optimization.

The segmenter is based on the Markov topology as depicted in the Figure 1. Each state of this Markov network is associated with one of the three broad audio type segmentation classes *noise*, *speech*, and *music*, respectively. Parameter re-estimation of the Gaussian mixture densities was performed using both a ML criterion and the MMI criterion.

Comparative experiments were performed on a German broadcast news corpus. Both segmenters use 16 cepstral coefficients with the 16 first derivatives and the second derivative of the energy. Table 1 shows the segmentation results. Compared to the ML based training the discriminative approach improved the segmentation purity from 97.57% to 98.84%. The amount of lost speech could be reduced from 1.58% to 0.89%. Details can be

Table 1: *Comparison of two segmentation approaches on the test set of the Report corpus.*

approach	speech lost [%]	purity [%]	avg. segment length	words cut [%]
ML	1.58	97.57	21.20s	0.17
MMI	0.89	98.84	26.30s	0.12

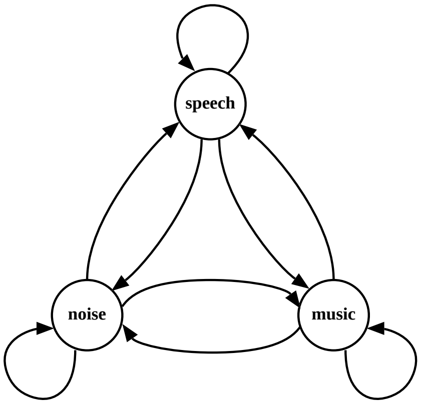

Figure 1: *Ergodic Markov topology for speech-noise-music segmentation.*

found in [19].

2.2 Model Selection Criteria for Acoustic Segmentation on Italian Broadcast News

In recent years, several algorithms for acoustic segmentation have been presented which use a statistical decision criterion to detect spectral changes (SCs) within the feature space of the signal. Assuming that data are generated by a Gaussian process, SCs are detected within a sliding window through a model selection method. The most likely SC is tested by comparing two hypotheses: (i) the window contains data generated by the same distribution; (ii) the left and right semi-windows, with respect to the SC point, contain data drawn by two different distributions.

The test is performed with a likelihood ratio that, besides the maximum likelihood of each hypothesis, takes into account the different “size” of the corresponding models. Usually, the Bayesian Information Criterion (BIC) [25] is applied to select the simplest and best fitting model. This work[1] reviews alternative model selection criteria and presents comparative experiments both on synthetic data and real audio data.

[1] This document is an excerpt from [6].

2.2.1 Segmentation with Transition Detection

Acoustic segmentation can be seen as a particular instance of the more general problem of partitioning data into distinct homogeneous regions [2]. The data partitioning problem arises in all applications which require to partition data into chunks, e.g. image processing, data mining, text processing, etc.

The problem can be formulated as follows. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n$ be an ordered sample of data in the $\Re^d$ space. We assume that the data are generated by a Gaussian process with at most c transitions. The problem of segmentation is that of detecting all the transition points in the data set. The general problem can be approached, without loss of generality, by first considering the simplest case $c = 1$.

Single Transition Detection

The search of one potential transition point goes through the definition of n different statistical models:

- $n-1$ two-segment models M_t ($t = 1, \ldots, n-1$), each of them assuming:

$$\boldsymbol{x}_1 \ldots \boldsymbol{x}_t \quad \sim_{iid} \quad N_d(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1) \tag{1}$$

$$\boldsymbol{x}_{t+1} \ldots \boldsymbol{x}_n \quad \sim_{iid} \quad N_d(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2) \tag{2}$$

- one single-segment model M_n which assumes:

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n \sim N_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma). \tag{3}$$

The basic idea is to choose the model ($M_t : t = 1, \ldots, n$) that better fits the observations. The application of the maximum likelihood principle would however invariably lead to choosing one of the two-segment models, and hence to hypothesize a break point at some $t = 1 \ldots n-1$, as they have a higher number of free parameters than the one-segment model. In order to take into account the notion of "dimension" of the model, the following extension to the maximum likelihood principle was first proposed by Akaike [1]. The AIC (Akaike's Information Criterion) suggests to maximize the likelihood for each model i separately, obtaining say $L_i = L_i(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n)$, and then choose the model for which $logL_i - k_i$ is largest, where k_i is the dimension of the model.

Computations

Given a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n \sim_{iid} N_d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ the likelihood function achieves the maximum value [26]:

$$L(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n) = (2\pi)^{-\frac{nd}{2}} | \, \hat{\Sigma} \, |^{-\frac{n}{2}} e^{-nd/2} \tag{4}$$

at $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$, where $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$, the sample mean, and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \tag{5}$$

the maximum-likelihood estimate of the covariance matrix. The number of free parameters of a multivariate normal distribution is equal to the dimension of the mean plus the number of variances and covariances to be estimated. For a full covariance matrix it is:

$$k = d + d\frac{(d+1)}{2} \ . \tag{6}$$

Decision Rule

Several model selection criteria have been proposed in the literature that can be applied to Akaike's framework of model selection. In general, each criterion proposes a penalty function P that takes into account the model dimension.

By computing the likelihood function of each model, the following decision rule can be derived. Look for the best two-segment model for the data:

$$logL_{t'} - P_{t'} = \max_{t=1,\ldots,n-1} -\frac{t}{2}log \mid \hat{\Sigma}_1 \mid - \frac{n-t}{2}log \mid \hat{\Sigma}_2 \mid - P_t \tag{7}$$

then, take the one-segment model function:

$$\log L_n - P_n = -\frac{n}{2} \log \mid \hat{\Sigma} \mid - P_n \tag{8}$$

and choose to segment the data at point t' if and only if:

$$(\log L_{t'} - \log L_n) - (P_{t'} - P_n) > 0. \tag{9}$$

In the experimental part it will be shown that performance of the rule can be tuned by replacing the zero threshold with a value θ to be empirically estimated.

Multiple Transition Detection

The extension of the method to an arbitrary large number of potential segments requires considering a number of competing models that combinatorially grows with n and c. In general, application dependent simplifications are introduced to reduce the complexity of the problem. For the acoustic segmentation, the audio signal can be segmented through a sliding window. By keeping the window size sufficiently large to reliably apply the method, and sufficiently short to avoid multiple transitions, a segmentation algorithm is proposed in [5] that relies on the basic $c = 1$ case. The main idea is to have a shifting variable-size window in which a SC can be hypothesized according to (9).

2.2.2 Model Selection Criteria

Several model selection criteria have been proposed starting from the early '70s. As mentioned before, the seminal work of Akaike tried to extend the maximum likelihood principle

with a term that estimates the dimension or complexity of the considered statistical model. Refinements to the Akaike's Information Criterion (AIC) were proposed by Schwarz [25], with the Bayesian Information Criterion (BIC), and by Bozdogan [3], with the Consistent AIC (CAIC), and the Consistent AIC with Fisher information (CAICF). By following an information and coding theory approach to statistical modelling and stochastic complexity, Rissanen [24] and Wallace and Friedman [29] proposed in the '80s two different criteria, respectively called Minimum Description Length (MDL) and Minimum Message Length (MML).
Without going into the details of each method which would require too much space, the penalty terms derived by each of the mentioned criteria are given in Table 2.

Name	Author	Year	Reference	Penalty
AIC	Akaike	1972	[1]	k
BIC	Schwarz	1978	[25]	$\frac{k}{2}\log n$
CAIC	Bozdogan	1987	[3]	$\frac{k}{2}\log n + \frac{k}{2}$
CAICF	Bozdogan	1987	[3]	$k + \frac{k}{2}\log n + \frac{1}{2}\log \mid I(\theta) \mid$
MDL	Rissanen	1987	[24]	$\frac{k}{2}\log n + (\frac{k}{2}+1)\log(k+2)$
MML	Wallace & Freeman	1987	[29]	$\frac{d}{2}(1+\log \kappa_d) + \frac{1}{2}\log \mid I(\theta) \mid$
Notation:				
k	Number of free parameters in the model.			
n	Size of the data sample.			
d	Dimension of the data space.			
$I(\theta)$	Fisher Information matrix of the model.			
κ_d	Constant of the optimal d-dimensional quantizing lattice [9].			

Table 2: Model Dimension Estimates.

For the sake of comparison, a version of the Hotelling's T^2 test and the maximum likelihood method are also considered.

Hotelling's Test

The Hotelling's T^2 (T2) test [28] computes the maximum likelihood estimate of a changing point of the mean in the sample by:

$$\begin{aligned} t' &= \arg \max_{t=1,\ldots,n-1} T_t^2 \qquad (10) \\ &= \arg \max_{t=1,\ldots,n-1} \frac{t(n-t)}{n-2} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \end{aligned}$$

where S_p is the pooled variance:

$$S_p = \frac{1}{n-2}(t\hat{\Sigma}_1 + (n-t)\hat{\Sigma}_2) \tag{11}$$

and $(\bar{\boldsymbol{x}}_1, \hat{\Sigma}_1)$ and $(\bar{\boldsymbol{x}}_2, \hat{\Sigma}_2)$ are, respectively, the sample means and ML covariance estimates on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_n$. The hypothesis of a changing point can again be accepted with a confidence level $(1-\alpha)$ if:

$$\frac{n-d-1}{(n-2)d} T_{t'}^2 \geq F_{d,n-d-1,\alpha} \tag{12}$$

where $F_{d,n-d-1,\alpha}$ is upper $\alpha 100\%$ point of the F-distribution with (d,n-d-1) degrees of freedom.

Maximum Likelihood Test

The Maximum Likelihood (ML) criterion corresponds to a model selection criterion with a zero penalty function. Hence, a SC is detected if the two-segment model fits the data better than the single-segment model.

2.2.3 Evaluation Metrics

Segmenting an audio stream, like a broadcast news program, requires in general to detect spectral changes regarding:

- acoustic sources, i.e. female/male speech, music
- acoustic channels, i.e. wide/narrow band.

According to [18], performance of automatic SC detection should be calculated with respect to a set of target SCs. To each target SC there is usually associated a time interval $[S_{SC}, E_{SC}]$, rather than as a single point. This because silence or other non-speech events may occur between changes. Tolerances in detecting SCs can be introduced by extending such intervals. Hence, an hypothesized SC is considered correct if it falls inside one of the augmented target intervals $[S_{SC} - tol, E_{SC} + tol]$, where tol is the admitted tolerance. For comparing target and hypothesized SCs, one can adopt the recall and precision measures:

$$\text{recall} = \frac{a}{a+c} \times 100 \qquad \text{precision} = \frac{a}{a+b} \times 100 \tag{13}$$

where a is the number of hypothesized SCs that fall inside the target SC intervals, b is the number of hypothesized SCs that do not fall inside any target SC interval, and c is the number of target SC intervals which no hypothesized SC falls inside.

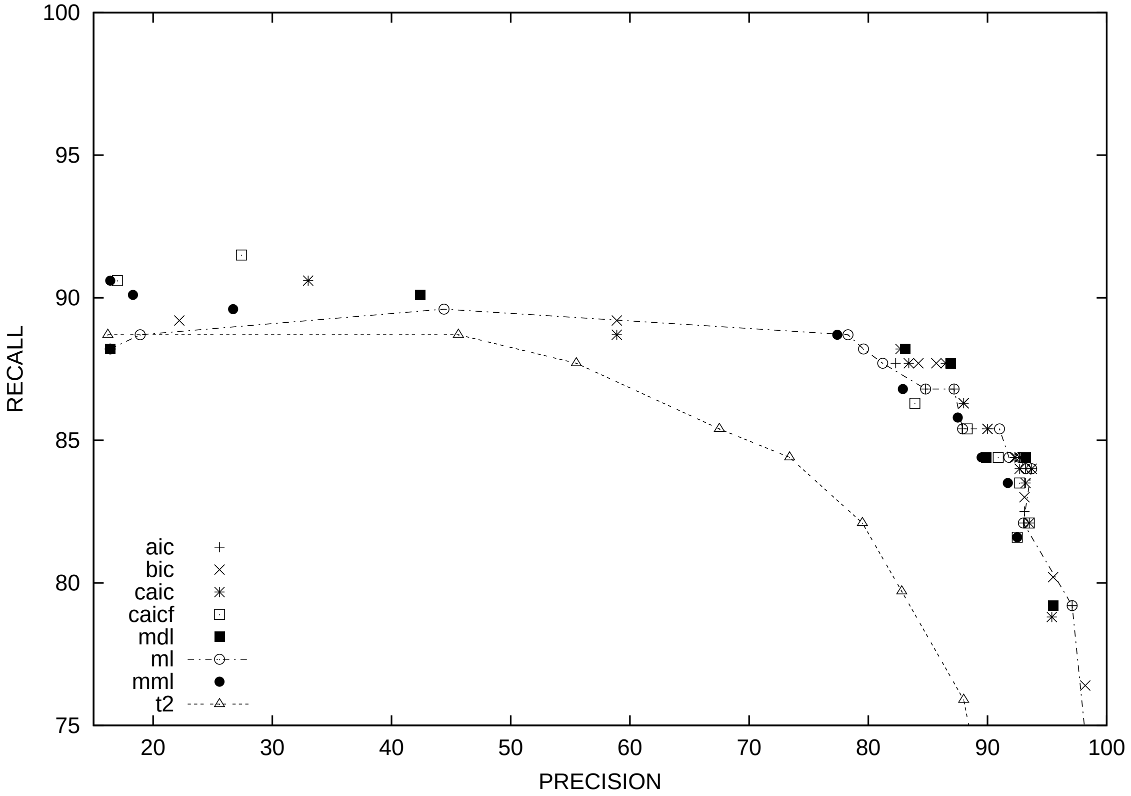

Figure 2: Precision vs. recall curves by different methods on the SC detection task.

2.2.4 Experiments

Experiments with all the segmentation criteria were performed on audio data coming from the Italian Broadcast News (IBNC) database, developed at ITC-irst [11] presented in section 3.1.1.

Multivariate observations of dimension 13 were used, i.e. 12 mel-scaled cepstral coefficients and the log-energy. SCs detections was performed by using a tolerance value of 500ms. In order to compute a precision/recall operating curve of each method, an empirical threshold was introduced in the decision criteria (9) and (12). In fact, the threshold can be seen as an empirically estimated additional penalty to the method. Different values of the threshold were tested and the resulting precision/recall statistics were computed.

Precision vs. recall points of each method are shown in Figure 2. As a reference, complete curves are plotted for the ML and T2 methods. The left most points of all the model selection criteria correspond to setting the threshold to the original value, i.e. zero. By looking at Figure 2 the following can be observed:

- straightforward application of the methods on audio data provides high recall but very low precision;

- by suitably tuning the threshold value, on each single method, much better performance can be achieved;
- optimal values of the threshold make all methods, with the exception of T2, perform comparably well;
- T2 performs significantly worse than all other methods. Moreover, no improvement was achieved even by using a “universal” pooled variance estimated as suggested in [30];
- BIC, CAIC, and MDL confirm to be among the best performing methods;
- the pure empirically tuned ML method performs as well as the best model selection methods;

In summary, application of any method on real audio data requires introducing an empirical threshold on the decision criterion. Tuning the threshold on each method permits us to achieve significantly better retrieval performance. Of the several model selection methods for acoustic segmentation presented and tested, almost all considered methods reached very similar optimal performance.

3 Speaker Tracking and Acoustic Environment Labelling

Speaker tracking is the process of following who says something in an audio stream. It has many applications ranging from identifying speakers specifically, since speaker identity is an important meta-data for building digital libraries, to pooling data from the same speaker to increase the performance of speaker-adaptive recognition systems. Tracking can broadly be divided into two problems: (i) locating the points of speaker change (segmentation); (ii) identifying the speaker in each segment (labelling/classification).

Speaker tracking in broadcast news is a difficult task due to the following reasons: (i) the number of different speakers can be large; (ii) the chance of encountering a new speaker with respect to previous recordings is very high; (iii) most of speech data comes from speakers whose identity is not much interesting for the digital archive, such as anchor men, interviewers and reporters, while important speakers to be identified, such as politicians, typically speak for short time intervals.

Techniques for segmenting the audio stream are presented in section 2. In the first part of this section techniques for classifying the audio stream are introduced and experimentally evaluated on Italian and American English Broadcast News data.

In addition to identifying the foreground speaker, the problem of automatically identifying the background environment is a task of importance for applications such as achieving noise robustness and environmental adaptation of speech recognition systems. This is a difficult problem because the background sources must be identified in the presence of the foreground speech signal. The feasibility of this task is investigated for American English Broadcast News in section 3.3.

3.1 Speaker Tracking on Italian Broadcast News

In this section, the problem of automatically identifying speakers in the Italian Broadcast News Corpus (IBNC) is investigated. First, statistics of interest for the speaker tracking problem are reported, for the IBNC database detailed in section 3.1.1. The experiments are carried out on the ITC-irst collected speech corpus of radio broadcast news in the Italian language [11].

3.1.1 The Italian Broadcast News Corpus

ITC-irst collected, under a contract with ELRA/ELDA, a speech corpus of radio broadcast news in the Italian language [11]. The corpus, called IBNC (*Italian Broadcast News Corpus*), consists of 150 recordings, for a total of about 30 hours, covering radio news of several years. Data were provided by RAI, the major national broadcaster. The corpus presents variations of topics, speakers, channel band (i.e. studio versus telephone), speaking mode (i.e. spontaneous versus planned), etc. It has been manually transcribed, segmented and labelled. Speaker gender and, when possible, identity are also annotated.

The test set consists of six radio news programs (about 75 minutes of audio signal) that were selected as a representative sample of the whole corpus, with respect to all the issues concerning automatic broadcast news transcription [4]. Table 3 reports statistics on the test set regarding segments. A segment is defined as a contiguous portion of audio signal, homogeneous in terms of acoustic source and channel. The set contains a total of 212 SCs (218 segments distributed among six news programs).

	#	average duration (*s*)
music segments	17	2.0
speech segments	201	22.3

Table 3: Statistics of segments in the test set.

By naming speakers with labels that also take into account the acoustic conditions, the total number of speaker labels is 1072, which cover 94% of the audio data available. The average number of named speakers in each program is 13.5, with a maximum of 50: this shows that the number of different speakers occurring in a recording can be high.

The chance of encountering a new speaker with respect to previous recordings can be inferred from Figure 3, where it is shown that the number of new speakers occurring in a program remains constant even after tens of hours of recordings. However, most of speech data is uttered by few speakers. Figure 4 shows that less than 9% of speakers (all of them are reporters) uttered half of the speech data in the corpus.

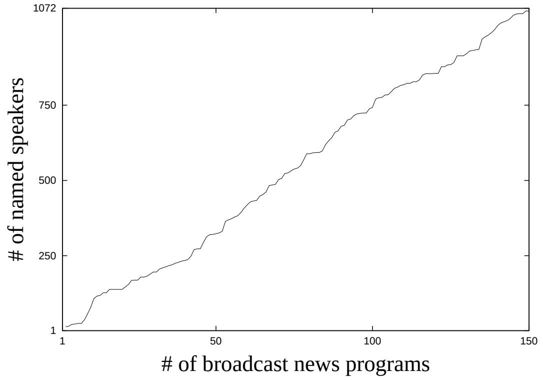

Figure 3: *Number of named speakers as a function of the number of news programs.*

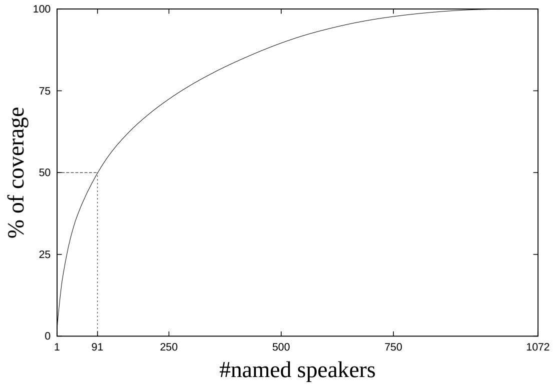

Figure 4: *Coverage (%) of speech data from named speakers as a function of the number of speakers, ordered in terms of the quantity of speech material available for each of them.*

For speaker tracking evaluation, six radio news programs (about 75 minutes of audio signal) were selected as a representative sample of the whole corpus. The rest of the IBNC was used for training purposes. The test set contains speech of 72 speakers; 63 of them have known identity. Among the 63 named speakers, only 37 utter more than 20 seconds of speech in the training material: these are the speakers who we aim at automatically tracking. For speakers who spoke in various acoustic conditions, e.g. with and without background music, a specific model for each condition has been trained. In Table 4, some statistics regarding these speakers in the training and test sets are reported.

The rest of the test data (about 33 minutes) is either speech uttered by speakers not modelled or non-speech. In order to cover it, a set of generic audio classes are needed; they have been modelled by using data selected from the training material not used for speaker modelling, for a total of 545 minutes. Table 5 contains details on data of the main

seconds available for each speaker	training	test
minimum	21.7	<1
average	400.4	68.0
maximum	2700.1	339.7
total	14815.6	2517.2

Table 4: Available data for the 37 test speakers.

of these classes. (The narrow-band classes include both telephone and noisy speech.) The generic classes, considered as a whole, represent the model of the world outside the known speakers.

class	training	test
	(seconds)	
wideband female	3682.2	78.0
wideband male	11298.9	598.9
narrow-band female	1571.0	199.4
narrow-band male	11801.1	910.9
music + female speech	702.3	64.7
music + male speech	1158.6	100.7
music	1426.9	27.1
other (silence, noise...)	1076.1	27.6
total	32717.0	2007.4

Table 5: Training/test sets statistics on generic classes.

3.1.2 Speaker Tracking Methods

Multivariate observations derive from a short time spectral analysis, performed over 20 *ms* Hamming windows at a rate of 10 *ms*. For every window, 12 Mel scaled Cepstral coefficients, the log-energy and their first and second order time derivatives are evaluated.

Gaussians mixture models [22] are employed to model each test speaker and each generic audio class. Emission probability densities consist of mixtures of multi-variate Gaussian components having diagonal covariance matrices. Different numbers of components were considered, i.e. 16, 32, 64 and 128, in order to evaluate the impact of more refined modelling on classification performance.

Segmenting an audio stream means to detect the time indices corresponding to changes in the nature of audio, in order to isolate segments that are homogeneous in terms of bandwidth and speaker. Our technique bases segmentation on a statistical model selection criterion, by applying the Bayesian Information Criterion (BIC) [25, 8, 5, 6]. According to the BIC, the decision of hypothesizing or not a change in a particular time index can be based on a threshold λ, that determines the sensitivity of the method, i.e. the lower λ is, the higher the number of hypothesized changes will be. This work is presented in section 2.2.

Once the acoustic segmentation stage outputs a sequence of acoustically homogeneous segments, each of them has to be assigned to one of the known speakers or to one generic class.

Three different classification techniques have been tested; they are briefly introduced in the following.

1. **The `ML` technique**: if the automatic segmentation is assumed to be reliable, each segment can be classified with the class giving the maximum likelihood. From here on, we refer to this technique with `ML`. The use of priors on classes, that would make the classification Bayesian, did not favorably impact on performance. Hence, equally distributed priors were assumed.

2. **The `loop` technique**: if we consider that the automatic segmentation can fail, as in fact it happens, we can classify not the whole segments, but single observations, in order to introduce segment boundaries not detected by BIC. Possibly, BIC can also insert spurious changes, but since we can tune the segmentation algorithm in such a way that deletions result more frequently than insertions, in this work we focused on the problem of boundary deletions.

   In order to classify each observation, a Viterbi decoding can be performed on a search space defined by the simple loop-based network of Figure 5. Once the classifier has labelled the current observation with a certain class, it will remain in that class or instantiate another class with probabilities defined by the factor α. The lower the α is, the more penalized the instantiation of a new class will result, i.e. the lower the number of new changes introduced by this stage will be. From here on, we refer to this technique with the term `loop`.

3. **The `hierarchy` technique**: a hierarchical combination of the two above techniques has also been tested. Since, on the average, there is much more training data for generic audio classes than for specific speakers, it appears convenient to split the classification into two stages. In the first stage, each segment is classified into one of the generic audio classes through the `loop` technique, that also allows recovering boundaries not detected by the BIC. The second stage is then performed only on segments assumed to contain speech data; the `ML` classifier is employed to identify the speaker on a search space restricted to speakers that belong to the class hypothesized during the first stage. The rejection is allowed, since the generic model of that class

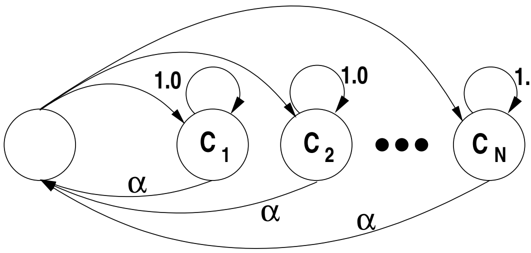

Figure 5: Network for the Viterbi-based classification algorithm.

is added to the restricted search space, too. From here on, this technique will be referred to as `hierarchy`.

3.1.3 Experiments

According to [12, 13], the classification performance is evaluated by computing the frame classification accuracy (FCA), that is the percentage of frames classified in the correct class out of the total number of frames.

Since this work focuses on the speaker tracking problem, the classes taken into account for evaluation are the 37 named speakers plus one class representing the rest of the world. This means that misclassifications inter-generic classes are not considered.

For the best result obtained, details are given in terms of miss and false alarm probability too, both for each speaker and on the average.

As a reference, we first run the automatic classification on the manually segmented test set. In this way, errors that can occur in the detection of audio changes are not considered. In Table 6, performance of the `ML` classifier are reported.

It is worth noticing that the use of more Gaussian components for acoustic modelling does not necessarily ensure better results. It seems that mixtures with more than 32 Gaussians are not well trained on average, given the training set employed in this work.

Figure 6 shows four plots, one for each number of Gaussian components employed in acoustic modelling.

Each plot includes three curves, corresponding to performance of the three tested classifiers: `ML`, `loop` and `hierarchy`. The αs of the two latter techniques were set to the

technique	#Gaussians			
	16	32	64	128
ML	79.0	79.8	78.2	78.2

Table 6: Performance of the ML technique in the classification of manual segments.

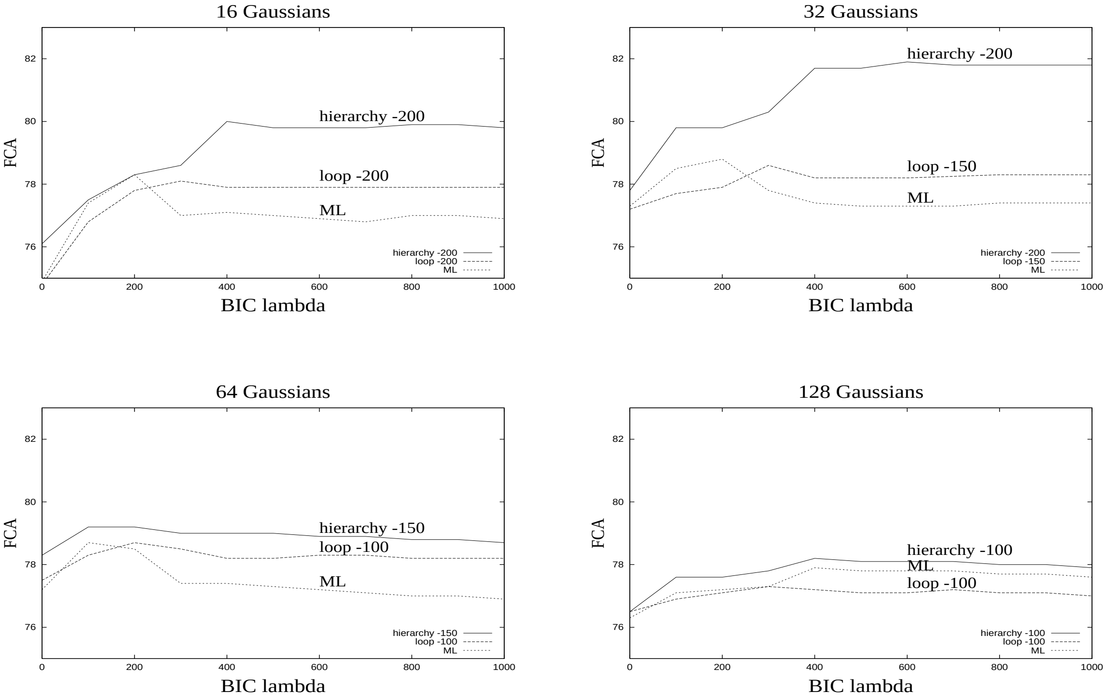

Figure 6: Results of Speaker Tracking experiments.

values that give the best performance. The curves are functions of the BIC threshold λ; the set of λ values tested ensures a large variability in BIC outputs, ranging from a lot of insertions of spurious boundaries ($\lambda = 0$) to the hypothesizing of only very reliable changes ($\lambda \geq 800$).

Looking at the experimental results, the following main observations can be done:

- The best performance are obtained by employing 32 Gaussians. The `ML` classifier gives 78.8% FCA as its best result; that is, the degradation of performance due to the automatic segmentation can be kept lower than 1.3% relative.
- The curves that have a protuberance for a certain interval of λ values, show that an optimal working point exists, where segmenter and classifier integrate their respective work in an effective way. However, it has to be noted that such protuberances are less evident, or do not exist at all, for `hierarchy` curves; typically, `hierarchy` curves have a monotone increase that ends when a sort of saturation is reached. In particular, this happens for the `hierarchy`-32 Gaussians plot, that includes the best performance of these experiments (81.9% FCA).
- The `hierarchy` classifier performs significantly better than the other tested techniques, and its best performance are kept for a large range of λ BIC values, making not critical the integration of BIC segmentation in the speaker recognition system.
- The best performance obtained by using the `hierarchy` technique (81.9% FCA) is even better than the FCA given by the `ML` classifier on manual segments. This is due to the fact that manual segments misclassified by the `ML` classifier may be split in shorter segments by the `hierarchy` processing, and it is sufficient that some of them are correctly classified to explain the result.

Finally, Table 7 shows details of the best experiment (81.9% FCA) for each speaker and for the world model that groups together all the generic classes. In this case, also miss and false alarm probabilities are given.

The weighted mean of miss probability rates of named speakers is 26.1%; it is worth noticing that most of the missing speech of named speakers is in fact classified into generic classes, and not into wrong speakers, since the false alarm probability of the world model (24.8%) is almost double of that of named speakers (12.8%).

Such errors would involve the introduction of wrong meta-data into the digital library, but the construction of the library may not be affected by them. In fact, the preprocessing described in this work is typically used in order to isolate the portions of speech inside the audio stream, to select the specialized acoustic models to be employed during the speech decoding, and to pool data for adaptation purposes. In such a case, important metrics for assessing the quality of the preprocessing are the quantity of speech lost, that of non-speech supplied to the recognizer, and gender and bandwidth error rates. In our best experiment, the following values have been computed:

class/speaker	F/M	test/training		miss	FA
		size (s)		probability (%)	
world model class	–	33.5m	545.3m	8.2	24.8
P. M.	M	50.3	71.6	0.0	0.6
N. Z.	F	84.1	57.3	0.0	0.8
C. A.	M	42.9	299.7	0.0	1.9
F. S.	M	55.6	173.1	0.1	0.7
C. C.	M	63.0	130.1	0.4	0.0
F. C.	M	71.3	262.7	0.4	0.0
R. P.	M	104.8	710.2	0.4	0.0
S. M.	F	67.8	231.0	0.4	0.8
F. F.	M	38.4	201.9	0.7	1.6
B. R.	M	20.7	173.9	0.9	106.1
V. A.	M	12.1	27.7	1.2	0.9
P. A.	F	100.8	610.1	1.2	2.7
R. V.	M	58.8	146.5	1.4	0.0
R. G.	M	52.2	1076.2	1.6	9.7
N. A.	F	44.3	85.6	2.0	3.9
L. N.	F	194.3	306.1	2.1	0.3
V. M.	F	94.2	2103.6	2.4	4.4
G. G.	M	15.5	750.8	3.1	318.1
F. C.	M	45.8	62.7	3.2	0.5
M. V. B.	F	339.7	680.6	4.2	10.8
P. M.	M	27.1	76.0	5.8	2.3
T. L.	F	161.7	2700.1	21.4	24.3
P. L.	M	243.6	1581.3	27.2	0.2
M. L.	F	6.8	516.6	72.6	107.3
A. C.	M	2.6	24.5	100.0	0.0
B. P.	M	47.4	47.0	100.0	0.0
D. G.	F	124.9	153.9	100.0	0.0
E. L. L.	M	51.3	24.3	100.0	0.0
E. R.	M	66.1	94.9	100.0	0.0
F. B.	M	7.3	50.5	100.0	0.0
F. B.	M	86.2	21.7	100.0	0.0
F. R. C.	F	0.4	265.8	100.0	0.0
N. B.	M	43.3	41.3	100.0	0.0
O. L. S.	M	17.7	238.6	100.0	0.0
V. C.	M	31.4	97.5	100.0	0.0
P. A.	M	42.8	465.1	100.0	155.0
G. D. C.	M	0.01	255.1	100.0	10^6
total sizes/weighted means		42.0m	246.9m	26.1	12.8

Table 7: Details of the best experiment (`hierarchy`, 32 Gauss., $\lambda_{BIC} = 600, \alpha = -200$) in terms of miss and false alarm (FA) probabilities.

speech lost	1.0%
non-speech decoded	0.2% (wrt speech) 12.8% (wrt non-speech)
gender error rate	2.4%
bandwidth error rate	4.5%

Table 8: Overall System Performance

Focus	Description
F0	baseline broadcast speech (clean,planned)
F1	spontaneous broadcast speech (clean)
F2	low fidelity speech (wideband/narrowband)
F3	speech in the presence of background music
F4	speech under degraded acoustic conditions
F5	non-native speakers (clean,planned)
FX	all other speech (e.g. spontaneous non-native)

Table 9: American English Broadcast News focus conditions

These results are summarized in table 8 comparable to those reported in the literature elsewhere.

3.2 Speaker Tracking on American English Broadcast News

The experiments were based on using a 1024 component Gaussian Mixture Model which forms the Universal Background Model (UBM). MAP adaptation of the means only of the UBM is performed with all the available speech data for each speaker. Thereby 50 speaker specific models were generated. The speakers are selected by simply extracting the most frequent (by number of frames) 50 speakers in the training corpus.

The experiments on American English Broadcast News is carried out on the training and development data corpus. Features used are 13 PLP features and normalized log energy, their delta and acceleration coefficients.

3.2.1 The American English Broadcast News Corpus

The US Broadcast News corpus (US-BN data) acoustic data sets of 1996, recorded and labelled by the Linguistic Data Consortium (LDC), consist of a collection of American television and radio shows. Two of the released data sets are used for the experiments in this work. [36]

The training data corpus (referred to as BNtrain96) consists of about 35 hours of transcribed data from the following collection of shows transmitted prior to June 30, 1996: ABC Nightline, ABC World News Now, ABC World News Tonight, CNN Early Edition, CNN Early Primetime News, CNN Headline News, CNN Primetime News, CNN The World Today, CSPAN Washington Journal, NPR All Things Considered, and NPR Marketplace.

The development test data corpus (referred to as BNdev96) consist of 6 complete shows broadcast in July 1996: ABC Prime Time, CNN World View, CSPAN Washington Journal, NPR Marketplace, NPR Morning Edition and NPR The World. A hand-partitioned segmentation is provided with the development data set which contain segments that are homogeneous with respect to the foreground speech and background environmental conditions.

The transcriptions provide meta-data for both data sets regarding the channel condition, an acoustic environment analysis and a ”focus condition”. The channel condition is marked as Telephone/Wideband. The background analysis is provided by three noise condition variables - background speaker, background music and other background noise - each marked as High/Low/Clear. Seven focus condition are provided as in Table 9.

The distribution of data in the US Broadcast News corpus (US-BN data) is very different to the distribution associated with the IBNC. Looking at the top 50 speakers from the training data only covers about 36% of the training data. In addition only 2 of these speakers are present in the test data covering about 15% of the frames of the test data (84 of the 553 automatic segments). In contrast to standard speaker identification data there is a very large number of speakers in the test data that are required to be rejected.

The automatically partitioned segments [13] are labelled by their dominant speaker. The segmenter produces 553 segments as opposed to 448 homogeneous (w.r.t. the speaker and the environment) manual segments. There are only 70 dominant speakers as opposed to 77 true speakers, which denotes that automatic segmentation has allowed 7 speakers to be completely dominated by the others. Furthermore, 6 segments are produced with no foreground speaker, each of which is assigned a new *SIL* speaker.

3.2.2 Experimental Results

This work uses standard MAP based speaker identification techniques using MAP adaptation and a universal background model (UBM) [23].

The performance of the speaker identification task is graphed against the rejection threshold in figure 7. The x-axis of the graphs show the threshold of the likelihood given by the model identifying the segment, less the likelihood output of the UBM identifying the segment. The likelihoods considered are per segment and per frame for the respective graphs. The y-axis shows the proportion of the test frames/segments concerned to the total number of frames/segments. A correct recognition is an actual known speaker cor-

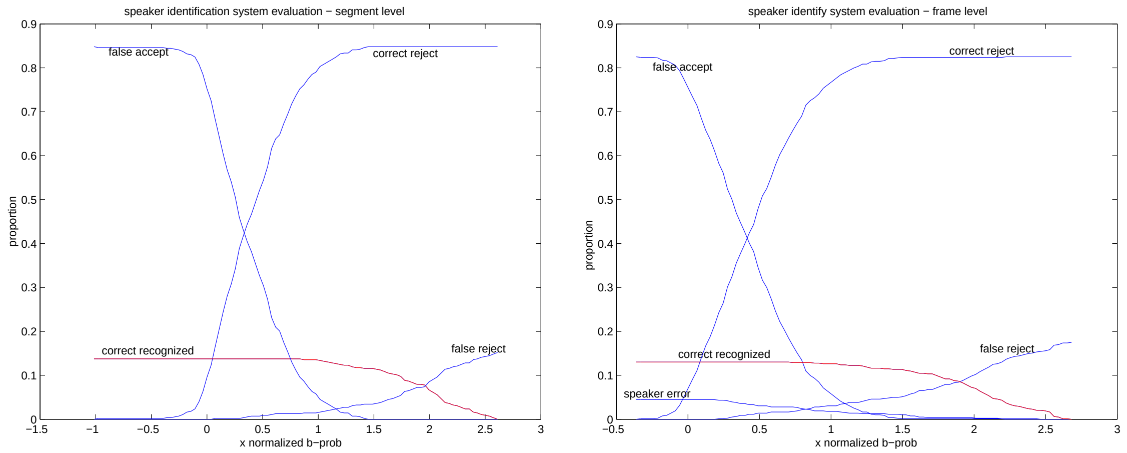

Figure 7: Speaker Tracking performance vs. rejection threshold graphs at dominant speaker labelled segment level (left) and frame level (right). The main curves giving the proportion of speakers correctly accepted, speaker correctly rejected as unknown, unseen speakers falsely accepted as known, and known speakers falsely rejected as unknown are shown. A speaker error is shown for frames incorrectly identified as a different speaker.

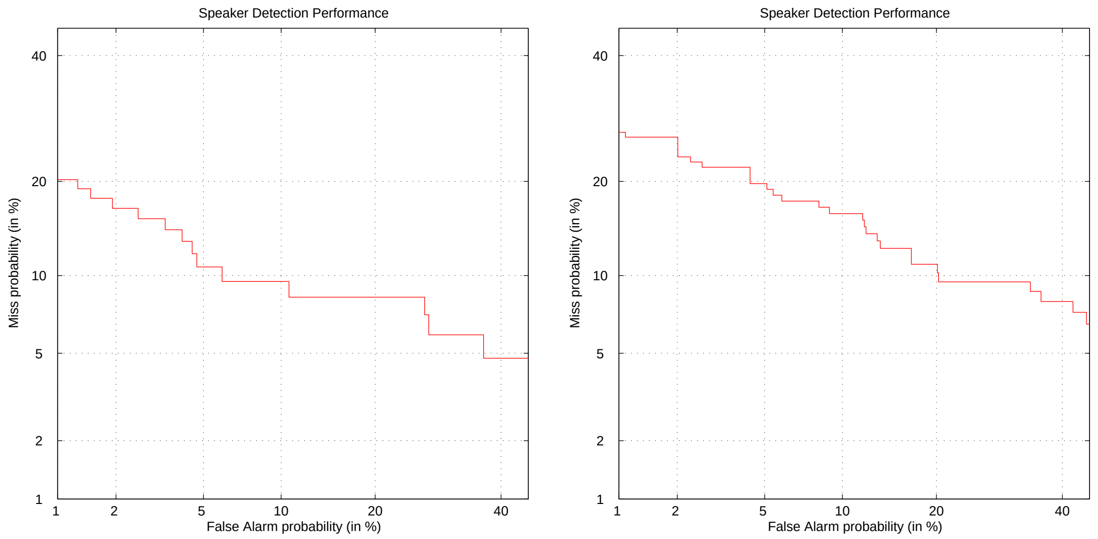

Figure 8: Speaker Tracking Detection graphs at dominant speaker labelled segment level (left) and frame level (right).

rectly recognized as such; a false acceptance is an imposter falsely recognized as one of the known 50 speakers (i.e. a false alarm). A correct rejection is an unseen speaker correctly rejected as 'unknown'; a false rejection is for one of the known speakers incorrectly rejected as 'unknown' (i.e. a miss). The speaker error is shown for frames of a known speaker identified as a different known speaker. The graph give the proportion of speakers that falls into each category with varying rejection ('recognition likelihood - the UBM likelihood') threshold.

The detection probability curves are shown in figure 8. In these curves the likelihood outputs for the 'missed' speakers are shown against the likelihood outputs of the 'false alarms', sorted in the decreasing order. The curves are close to linear, particularly the frame-based curve, which confirms that the likelihood distributions of known and unknown speakers are both normally distributed about their respective means. [20]

When the rejection threshold is set to EER (Equal Error Rate where the false acceptances are equal to the false rejections), at dominant speaker labelled segment level, of the 84 seen segments 70 (83.3%) are correctly identified; of the 469 unseen segments 457 (97.4%) are correctly rejected as 'unknown'.

3.2.3 What can be done with Unseen Speakers?

Particularly for US-BN data there is a large proportion of unseen speakers. Clustering techniques are employed as elaborated in section 4.3 to group together segments of the same speaker. This is done (i) approximating the number of required clusters assuming that the required number of important unknown speakers is not known (ii) assuming the required number of important speakers is known.

The clustering scheme that grows a tree of moderate depth with recombinations achieves a clustering of 100 clusters that shows an efficiency (BBN [0,1] scaled) of 0.657 at segment level and 0.661 at frame level. By marginally adjusting the clustering scheme (by increasing the threshold gain required to split nodes and the allowed recombinations) the number of clusters is brought close to the true number of unknown speakers of 71. This scheme produces 75 clusters with a (BBN [0,1] scaled) clustering efficiency of 0.659 at segment level and 0.692 at frame level.

3.3 Acoustic Environment Labelling on American English Broadcast News

In this section the task of identifying the background noise condition of cleaned and pre-segmented acoustic data, where speech remains in the foreground, is investigated. Such information on the acoustic environment can be used for environment adaptation and for achieving noise robustness of speech recognition systems. This acoustic environment markup is carried out according to three environment labelling schemes for the American English 1996 Hub-4 Broadcast News data.

3.3.1 Environment Labelling Schemes

The three environment labelling schemes considered are derived from three meta-data variables released with the 1996 Hub-4 Broadcast News transcriptions - background music, background speaker and other noise - each marked as Clear, High or Low.

Env2 scheme: In this scheme the environment has 2 attributes on whether there is any noise or no noise in the background.

Env3 scheme: According to this scheme the environment has 3 attributes on whether there is no noise, or whether the noise level is High or Low. According to the release notes some level of subjective human error in marking loudness levels may be expected, and this scheme may suffer from this low confidence in labelling.

Music scheme: This scheme does an analysis that gives priority to music where the environment is analyzed as clear, whether there music in the background or, where there is no music, contains any other noise. It is considered to investigate whether a clustering scheme is able to capture signature tunes or patterns in the music over other background noise. It is noted that 3 segments with low music and high other noise are labelled as music.

3.3.2 Experiments

Maximum Likelihood GMM classifiers are trained on the 1996 Hub-4 Broadcast News training (BNtrain96) data corpus and applied to recognize the 1996 automatic segments of Hub-4 Broadcast News development (BNdev96) data corpus, according to the three environmental labelling schemes. Two classifiers are built to identify the music and the env3 “loudness” schemes respectively. The env2 scheme results are derived from the env3 scheme. The classifiers, with 1024 mixtures for each class, are trained on the complete BNtrain96 data corpus; two separate telephone classifiers, with 1024 mixtures per class, are retrained from these on separated telephone data only. Increasing the number of mixture components beyond 1024 does not provide performance gains.

Features used are 13 PLP features and normalized log energy, their delta and acceleration coefficients. Using Cepstral Mean Normalization degraded environment detection performance.

The experiments are carried out to mark up the background environment conditions of the automatically generated segments. [13] The US Broadcast News corpus is described in section 3.2.1 and the effects of automatic segmentation are described in section 4.4. The automatic segments are classified according to bandwidth and recognized separately by the wideband and telephone classifiers. The recognition results are scored at dominant labelled segment level, where each segment is assigned the label of the dominant number of frames in it, and at frame level, where credit is assigned to all correctly recognized frames.

	env2	**env3**	**music**
Correctly Classified Segments	88.6%	81.9%	79.7%
Correctly Classified Frames	90.5%	84.4%	82.8%

Table 10: Recognition results of environmental classification for the 3 environmental schemes.

Segment Confusion	Classified Clear	Classified Noise
Labelled Clear	349	46
Labelled Noise	17	141
Frame Confusion	Classified Clear	Classified Noise
Labelled Clear	414364	35439
Labelled Noise	22893	141814

Table 11: Recognition confusion matrices for env2 labelling scheme

Test Data	Clear	Noise
Segment Distribution	71.5%	28.5%
Frame Distribution	73%	27%
Training Data	Clear	Noise
Segment Distribution	60%	40%
Frame Distribution	69%	31%

Table 12: Test data and Training data distributions for env2 labelling scheme

Segment Confusion	Classified Clear	Classified High Noise	Classified Low Noise
Labelled Clear	349	10	36
Labelled High Noise	2	23	15
Labelled Low Noise	15	22	81
Frame Confusion	Classified Clear	Classified High Noise	Classified Low Noise
Labelled Clear	414364	4707	30732
Labelled High Noise	2880	14043	15667
Labelled Low Noise	20013	21351	90753

Table 13: Recognition confusion matrices for env3 labelling scheme

Test Data	Clear	High Noise	Low Noise
Segment Distribution	71.5%	7%	21.5%
Frame Distribution	73%	5.5%	21.5
Training Data	Clear	High Noise	Low Noise
Segment Distribution	60%	4%	36%
Frame Distribution	69%	3%	28%

Table 14: Test data and Training data distributions for env3 labelling scheme

3.3.3 Environment Labelling Results

The recognition results are presented in Table 10. The recognition confusion matrices and data distributions are given for each labelling scheme in Tables 11 through 16.

The “noise presence” scheme, env2, has a 88% segment recognition rate; 88.4% of Clear labels and 89.2% of the Noise labels are correctly classified. Further investigation reveals evidence that the foreground signal effects the environmental classification, since a number of Clear segments misclassified as Noise belong to a particular speaker with a distinctively “hoarse” voice.

The “loudness” scheme, env3, has a 82% segment recognition rate; 88% of the Clear labels, 57.5% of the High Noise segments and 68.6% of the Low Noise segments are correctly classified. A factor contributing to the high proportion of misclassification in noise levels could be the low confidence in the marking of the noise levels.

There is a lower 79.7% recognition of the music scheme. While 85.8% of the Clear segments and 79.6% of the Other Noise labels are correctly identified, there is a high error of music segments being classified as Noise. It is noted that a number of segments with

Segment Confusion	Classified Clear	Classified Music	Classified Other Noise
Labelled Clear	339	11	45
Labelled Music	2	24	34
Labelled Other Noise	12	8	78
Frame Confusion	Classified Clear	Classified Music	Classified Other Noise
Labelled Clear	400334	13192	36277
Labelled Music	3853	22115	27919
Labelled Other Noise	15021	9547	86252

Table 15: Recognition confusion matrices for music labelling scheme

Test Data	Clear	Music	Other Noise
Segment Distribution	71.5%	10.5%	18%
Frame Distribution	73%	9%	18%
Training Data	Clear	Music	Other Noise
Segment Distribution	60%	10.3%	29.7%
Frame Distribution	69%	7%	24%

Table 16: Test data and Training data distributions for music labelling scheme

dominant label of Music, but which also contain other noise frames, are being classified as Noise, especially where the Noise level is high. In the two instances where low music and high other noise exists, the segments are being recognized as Noise. There is evidence that the classifier recognizes loud noise over music patterns in the segment.

4 Cluster Assessment

It often becomes necessary to be able to track a speaker through varied noise environments, for applications such as tracking a speaker through a telephone archive, and retrieving his speech from a Broadcast News database. It is also useful to successfully group together speakers in varied noise conditions for speaker and environmental adaptation.

In these applications where clustering is utilized, it is necessary to be able to evaluate what kind of clustering is carried out by a particular clustering scheme on a particular kind of data. This allows us to select the clustering scheme that provides a suitable clustering for the data, and shows what attributes of the data affect the clusters most. This can be evaluated for a given clustering scheme by labelling the clustering units according to varied possible attributes - such as speaker, gender, or speaker and noise condition for a speech database - and evaluating the labelling schemes according to some scoring criteria.

In this section the problem of evaluating clustering schemes with respect to varied labelling schemes is investigated. The metrics of evaluation are introduced and normalized for the joint evaluation of varied clustering and labelling schemes. The merits of the metrics are discussed. The techniques and the metrics are applied for (i) speaker-environment tracking for labelled homogeneous segments, and (ii) speaker turn assignment for unlabelled automatic segments, in American English Broadcast News.

4.1 The Clustering Problem

The clustering units are an S number of segments each with one of $N(1 \leq N \leq S)$ labels associated with it. The segments are distributed amongst the N labels according to some distribution and form the segment-label distribution vector $\mathbf{l} = (l_1, l_2, \dots, l_N)$ where l_i is the number of segments with label L_i. Given a fixed segment-label distribution vector, the clustering problem is one of assigning the segments to exactly N clusters such that all of the segments labelled L_i are in a single pure cluster labelled C_i. This ideal clustering can be represented by a diagonal matrix $\mathbf{M_0} = diag\{(l_1, l_2, \dots, l_N)\}$

However, a clustering algorithm may produce results that differ from the ideal in two ways. It may produce a number K of clusters that is not exactly N; it may distribute segments with the same label in different clusters and may include segments with different labels in the same cluster. This produces matrices that have more than or less than N columns, and have non-zero off-diagonal elements.

$$\begin{array}{cc}
 & \begin{array}{cccccc} C_1 & C_2 & \dots & \dots & \dots & C_K \end{array} \\
\begin{array}{c} L_1 \\ L_2 \\ \vdots \\ L_N \end{array} & \begin{pmatrix} s_{11} & s_{12} & \dots & \dots & \dots & s_{1K} \\ s_{21} & s_{22} & \dots & \dots & \dots & s_{2K} \\ \vdots & \vdots & & & & \vdots \\ s_{N1} & s_{N2} & \dots & \dots & \dots & s_{NK} \end{pmatrix}
\end{array}$$

$$l_i = \sum_{j=1}^{K} s_{ij} \quad , i = 1 \dots N; \qquad c_j = \sum_{i=1}^{N} s_{ij} \quad , j = 1 \dots K$$

A measuring scheme measures to which degree the resulting clustering matrix differs from the ideal clustering.

Secondly, it is important to conclude what a clustering result of a clustering scheme best represents. A clustering result may be analyzed with respect to different labelling schemes, where each scheme considers a different attribute or combination of attributes of the clustering unit. By changing labelling schemes, the segment-label distribution vector **l** and the number of labels N of the clustering matrix are now allowed to vary.

$$\begin{array}{rcl}
\mathbf{Scheme1}: & \mathbf{l} & = (l_1, l_2, \dots, l_N) \\
\mathbf{Scheme2}: & \hat{\mathbf{l}} & = (\hat{l}_1, \hat{l}_2, \dots, \hat{l}_N, \dots, \hat{l}_{\hat{N}}) \\
\sum_{i=1}^{N} l_i & = & \sum_{i=1}^{\hat{N}} \hat{l}_i \quad = \quad S \quad \text{(constant)}
\end{array}$$

A clustering metric is utilized in two ways. First, it is utilized to select, from a number of clustering schemes, the scheme that produces the best set of clusters for the considered fixed labelling scheme. The number of clusters K may vary from one scheme to another, and the measure must be normalized not to introduce an inherent bias with variation in the dynamic range.

Secondly, it is utilized for the evaluation of the result of a fixed clustering scheme with respect to varied labelling schemes. This would show what kind of clustering is carried out by the selected scheme on the particular kind of data. The measure of evaluation must also be normalized not to introduce inherent biases in the dynamic range with variation in the labelling scheme **l** or the number of labels N. This process of joint evaluation of clustering and labelling scheme allows us to select the clustering scheme that provides a suitable clustering for the data, and shows what attributes of the data affect the clusters most.

Evaluation can be carried out at two levels.

Segment based: the unit of evaluation is the unit weighted segment. An approximation is made that the segments are homogeneous and of approximately equal length.

Frame based: Information on variation in segment lengths must be captured for the successful comparison of clusters with a few large segments and clusters with many shorter segments. This is achieved by considering the frame to be the unit of evaluation by assigning each segment a weight equivalent to the number of frames in it. This problem can be represented by the clustering matrix $[w_{ij}]$ (where w_{ij} is the number of frames labelled i in cluster j), the corresponding frame-label distribution vector **l** with elements $l_i = \sum_{j=1}^{K} w_{ij}$, $c_j = \sum_{i=1}^{N} w_{ij}$, the perfect clustering matrix $diag\{\mathbf{l}\}$, and the total number of frames W.

Where the segments are not truly homogeneous, the simpler segment based evaluation can be carried out by approximating that the dominant label of the segment own all of the frames in the segment.

4.2 Clustering Metrics

A number of clustering metrics, such as entropy, variance, Gini and misclassification metrics, exist for determining the purity and efficiency of a clustering result.[10] This paper considers the Rand metric [14] and the BBN efficiency metric [27].

4.2.1 The Rand Metric

The Rand metric [14] of the clustering matrix **M** is defined as :

$$I_{RAND} = \frac{1}{2}\sum_{j=1}^{K} c_j^2 + \frac{1}{2}\sum_{i=1}^{N} l_i^2 - \sum_{i=1}^{N}\sum_{j=1}^{K} s_{ij}^2 \tag{14}$$

The I_{RAND} measure has value 0 for ideal clustering and a positive value that reflects the degree to which a **M** differs from the ideal M_0. To adapt the measure for the evaluation of varied labelling schemes, consider the metric result when the segment distribution amongst the clusters is random. Given a segment-label distribution vector **l**, using the uniform random assignment of the segments of each speaker to K clusters ($s_{ij} = l_i/K$), it can be shown that

$$random(I_{RAND}) = \frac{S^2}{2K} + \frac{(K-2)}{2K}\mathbf{l}'\mathbf{l} \tag{15}$$

The graph of Fig. 9 shows the *random(I_{RAND})* dynamic range variation with K for the speaker segment-label distribution of 1996 Hub-4 Broadcast News data. For the range of clusters generally produced by a clustering scheme a variation of about 1500 index points is shown. The second term of expression (15) also captures the variation introduced by variation of the labelling scheme. The $random(I_{RAND})$ readings as the labelling scheme varies are shown for four considered labelling schemes in the top section of Table 17.

The expressions for the frame-based metric is given by substituting s_{ij} with w_{ij} and S with the total number of frames W in (14) & (15), where **l** is now the frame-label distribution. In this expression, the term W^2 is very large and should make $random(I_{RAND})$ less sensitive to variations in K and **l**. The frame-based $random(I_{RAND})$ vs K graph is flat;

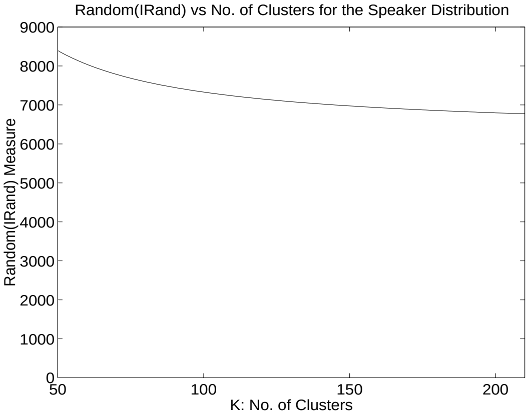

Figure 9: Random(I_{RAND}) vs K for 488 homogeneous segments with 77 speaker labels.

$random(I_{RAND})$ values for four different labelling schemes considered are shown in the top section of Table 18.

Normalized I_{RAND} is defined to enable comparison of clusters with varying K and segment/frame-label distributions:

$$Norm(I_{RAND}) \quad = \quad 1 - \frac{I_{RAND}}{random(I_{RAND})} \tag{16}$$

4.2.2 The BBN Metric

The BBN index [27] measures the efficiency of the clustering matrix **M** with respect to $\mathbf{M_0}$. It is defined in terms of p_j, the purity of cluster j and c_j the total number of segments in cluster j. The metric is already normalized and is appropriate for comparison of varying clustering and labelling schemes. Normalization is achieved in terms of $I_{BBN}(\mathbf{M_S})$ for singleton clustering with each segment in a separate cluster, $I_{BBN}(\mathbf{M_0})$ for perfect clustering and $I_{BBN}(\mathbf{M_1})$ for one cluster containing all segments, given the total number of segments S and the segment-label distribution vector **l**.

$$\begin{aligned}
p_j &= \sum_{i=1}^{N} \frac{s_{ij}^2}{c_j} && (17)\\
\eta_{BBN} &= \frac{I_{BBN}(\mathbf{M}) - I_{BBN}(\mathbf{M_S})}{I_{BBN}(\mathbf{M_0}) - I_{BBN}(\mathbf{M_S})} && (18)\\
I_{BBN}(\mathbf{M}) &= \sum_{j=1}^{K} c_j p_j - QK && (19)\\
I_{BBN}(\mathbf{M_0}) &= S - QN && (20)\\
I_{BBN}(\mathbf{M_1}) &= \frac{\mathbf{l'l}}{S} - Q && (21)
\end{aligned}$$

$$I_{BBN}(\mathbf{M_S}) \quad = \quad S(1-Q) \tag{22}$$

The frame-based BBN measure is defined by substituting w_{ij} for s_{ij} and W for S, and deriving **l** and c_j as appropriate for equations (17) through (21). However, the frame-based $I_{BBN}(\mathbf{M_S})$ must be experimentally estimated for the case where each cluster has a single segment in it. It is not appropriate to consider the case where each cluster has a single frame as the singleton measure, since this level of clustering is not achievable and causes measure insensitivity to clustering performance improvements.

Q is a user defined parameter representing the trade off between a few large clusters with many mixed labels, and many small clusters where labels may have more than 1 cluster associated with them. For the segment-based evaluation, Q is set to 0.5 to allow η_{BBN} to give a value in [-1,1] where 1 is achieved for perfect clustering, $I_{BBN}(\mathbf{M_1})$ achieves value -1, and negative values indicate a tendency to have large clusters with mixed labels. When Q achieves a critical value Q_{crit}, $I_{BBN}(\mathbf{M_1})$ achieves value 0, and η_{BBN} is normalized to give a value in [0,1]. Q_{crit} varies as the segment-label distribution varies (i.e. the labelling scheme changes). For the segment-based case

$$Q_{crit} \quad = \quad \frac{S^2 - \mathbf{l}'\mathbf{l}}{S(S-1)}$$

However, for frame-based evaluation Q_{crit} must be experimentally estimated to scale η_{BBNfr} to be in [0,1]. The scaling factor Q_{neg} must also be estimated experimentally for frame-based evaluation to scale η_{BBNfr} to be in [-1,1].

4.3 Cluster Assessment for Speaker-Environment Tracking on American English Broadcast News

The clustering metrics are utilized to evaluate three different clustering schemes with respect to four different labelling schemes. The data clustered are 488 segments of 1996 Hub-4 US Broadcast News Transcription development (BNdev96) data homogeneous with respect to the speaker and the environment. [13] The US-BN dataorpus is described in section 3.2.1.

4.3.1 Clustering Schemes

The segments are pre-classified into 4 categories according to gender and bandwidth and the 3 clustering schemes [15] are applied to each category. Top-down clustering is performed using the AHS distance measure between single Gaussian segment models. [16] [13] The 'adapt_c' scheme clusters by growing a clustering tree that terminates on a minimum occupancy count. The 'speaker1_c' scheme grows a tree until the gain from growth falls below a threshold, then recombines nodes that have a distance between the nodes that is less than twice the average distance within the node. The ‘speaker2_c’ scheme grows a larger tree that terminates on a larger gain growth threshold, and allows fewer recombination of nodes that have an inter-node distance less than the average within-node distance.

4.3.2 Labelling Schemes

Four different labelling schemes are considered [35]. The environment labels are the schemes detailed in section 3.4.2, and are combined with the speaker labels of the segments.
Spkr scheme: This scheme labels each segment with its speaker resulting in 77 labels.
Spkr-env2 scheme: This scheme gives a label to each speaker and environment condition combination; the environment is the ‘env2’ scheme marking noisy/clear conditions. There are 103 resulting labels.
Spkr-env3 scheme: This scheme gives a label to each speaker and environment condition combination; the environment is the ‘env3’ scheme marking high-noise/low-noise/clear conditions. There are 117 resulting labels.
Spkr-music scheme: This scheme gives a label to each speaker and environment condition combination; the environment is the ‘music’ scheme marking music/other-noise/clear conditions. There are 116 resulting labels.

4.3.3 Evaluation of Clustering Schemes

Tables 19 and 21 present the results of the metrics for clustering and labelling scheme evaluation. In the top section of the tables, the normalization metric $norm(I_{RAND})$, the current Rand metric and normalized Rand metric readings are shown. The two sections below show the [0,1] scaled and [-1,1] scaled BBN metric readings respectively. In each section the clustering schemes, given in bold along the left column, are evaluated according to the 4 labelling schemes shown. It is noted that according to the Rand measure, the lower the reading the better the ranking, while according to $norm(I_{RAND})$ the higher the reading the better the ranking.

Comparing along the columns for segment based clustering scheme evaluation for each fixed labelling, both Rand metrics and the BBN [0,1] metric agree that the ‘speaker1_c’ scheme is the best clustering scheme. However the BBN [-1,1] metric favors the ‘speaker2_c’ scheme albeit by a very small difference. However, in the frame based evaluation that is more sensitive to the length of the segments the BBN [-1,1] metric corrects itself, and agrees with the other metrics that the ‘speaker1_c’ scheme is the best clustering scheme for ‘spkr’, ‘spkr-env2’, and ‘spkr-env3’ labelling schemes. In particular for the ‘spkr’ and ‘spkr-env2’ labelling schemes, close observation of the clusters produced affirms the conclusion that the ‘speaker1_c’ produces the best clusters. The exception is the ‘spkr-music’ labelling scheme, where the BBN [-1,1] metric maintains that the ‘speaker2_c’ scheme is better. Hence it is difficult to draw a decisive conclusion for the ‘spkr-music’ labelling

	No.	**spkr**	**spkr**	**spkr**	**spkr-**
	Clust		**-env2**	**-env3**	**music**
No of Labels		77	103	117	116
Adapt_c:	81				
$rand(I_{RAND})$		*7580*	*5885*	*5385*	*5243*
I_{RAND}		*5376*	*4138*	*3759*	*3599**
$Norm(I_{RAND})$		0.291	0.297	0.302	0.314*
Speaker1_c:	92				
$rand(I_{RAND})$		*7423*	*5723*	*5221*	*5078*
I_{RAND}		*4286*	*3346*	*3167*	*2965**
$Norm(I_{RAND})$		**0.423***	0.416	0.393	0.416
Speaker2_c:	165				
$rand(I_{RAND})$		*6911*	*5194*	*4687*	*4543*
I_{RAND}		*4937*	*3665*	*3270*	*3106**
$Norm(I_{RAND})$		0.286	0.294	0.370*	0.316
Adapt_c:	81				
$\eta_{BBN}Q_{crit}$		0.646	0.537	0.511	0.516
Speaker1_c:	92				
$\eta_{BBN}Q_{crit}$		**0.707**	0.595	0.564	0.564
Speaker2_c:	165				
$\eta_{BBN}Q_{crit}$		0.616	0.511	0.484	0.496
Adapt_c:	81				
$\eta_{BBN}Q = 0.5$		0.336	0.055	-0.037	-0.026
Speaker1_c:	92				
$\eta_{BBN}Q = 0.5$		0.436	0.194	0.092	0.094
Speaker2_c:	165				
$\eta_{BBN}Q = 0.5$		**0.464**	0.206	0.123	0.147

Table 17: Segment based evaluation of clustering & labelling schemes for homogeneous data.

	No Cl	**spkr**	**spkr -env2**	**spkr -env3**	**spkr-music**
No of Labels		77	103	117	116
Adapt_c:	81				
$rand(I_{RAND})$		*10.56e9*	*9.01e9*	*8.65e9*	*8.64e9*
I_{RAND}		*6.38e9*	*5.17e9*	*4.89e9*	*4.86e9**
$Norm(I_{RAND})$		0.396	0.427	0.434	0.438*
Speaker1_c:	92				
$rand(I_{RAND})$		*10.03e9*	*8.78e9*	*8.41e9*	*8.40e9*
I_{RAND}		*4.36e9*	*3.42e9*	*3.39e9*	*3.29e9**
$Norm(I_{RAND})$		0.578	**0.610***	0.598	0.608
Speaker2_c:	165				
$rand(I_{RAND})$		*9.57e9*	*8.00e9*	*7.61e9*	*7.63e9*
I_{RAND}		*4.65e9*	*3.49e9*	*3.28e9*	*3.23e9**
$Norm(I_{RAND})$		0.514	0.564	0.570	0.576*
Adapt_c:	81				
$\eta_{BBN} Q_{crit}$		0.732	0.684	0.675	0.681
Speaker1_c:	92				
$\eta_{BBN} Q_{crit}$		**0.760**	0.715	0.699	0.702
Speaker2_c:	165				
$\eta_{BBN} Q_{crit}$		0.668	0.628	0.613	0.623
Adapt_c:	81				
$\eta_{BBN} Q_{neg}$		0.514	0.390	0.353	0.365
Speaker1_c:	92				
$\eta_{BBN} Q_{neg}$		**0.588**	0.468	0.418	0.425
Speaker2_c:	165				
$\eta_{BBN} Q_{neg}$		0.569	0.461	0.417	0.436

Table 18: Frame based evaluation of clustering & labelling schemes for homogeneous data.

scheme. This may be due to this being a poor labelling scheme.

It can also be concluded that normalization has not affected the ranking of the clustering schemes according to the Rand measures.

4.3.4 Evaluation of Labelling Schemes

Comparing along the rows for each fixed clustering scheme, both BBN metrics agree that the ‘spkr’ scheme is the most appropriate labelling for all clustering schemes. There is agreement between the BBN metrics, particularly at frame level, that the selected ’speaker1_c’ scheme carries out a ‘spkr-env2’ clustering of close efficiency to a ‘spkr’ based clustering. Close observation of the clusters produced by the selected ‘speaker1_c’ scheme shows that, while the clusters are primarily produced at speaker level, there is clear evidence that a secondary grouping of speakers according to a noisy/clear background is also being achieved. This confirms that, for this Broadcast News data clustering scheme, while the primary distances between segments are judged according to the foreground speaker signal, some influence could be attributed to background noise.

The best ranking labelling scheme according to the Rand metrics are marked by *. The negative effect of the bias on the current I_{RAND} metric is clearly demonstrated by strictly increasing ranking with decreasing $random(I_{RAND})$ readings. While normalized I_{RAND} metric is in somewhat better agreement with the observations and the conclusions, it is still unreliable for labelling scheme evaluation.

In summary it can be concluded that the Rand metrics are appropriate for ranking clustering schemes for a fixed labelling scheme. It can also be concluded that, while some improvement can be achieved by normalization of the Rand metric, the BBN efficiency metrics are preferred for the evaluation of labelling schemes. The experiments reveal that the clustering scheme which grows a tree of moderate depth with re-combinations, while achieving a frame-based speaker level clustering of high efficiency (0.760), also tracks the speakers through a noisy/clear environment with a relatively high efficiency of 0.715. This demonstrates that the background noise exerts a high level of influence on this application of speech data clustering.

4.4 Cluster Assessment for Assigning Speaker Turns on American English Broadcast News

In these experiments automatic segments formed by partitioning the stream of 1996 Hub-4 Broadcast News (BNdev96) speech data, are clustered to assign speaker turns. Its transcription is utilized to evaluate the success of this process for determining speaker turns.

The BNdev96 automatically partitioned segments [13] are labelled by their dominant speaker and environmental condition labels according to the labelling schemes detailed in section 4.3.2. The segmenter produces 553 segments as opposed to 448 homogeneous (w.r.t. the speaker and the environment) manual segments. There are only 70 dominant

	No.Clusters	**spkr**	**spkr-env2**	**spkr-env3**	**spkr-music**
No of Labels		71	93	101	105
Adapt_c scheme: $Norm(I_{RAND})$	106	0.248	0.265	0.270	0.270
Speaker1_c scheme: $Norm(I_{RAND})$	119	0.396	0.400	**0.407**	0.406
Speaker2_c scheme: $Norm(I_{RAND})$	151	0.359	0.377	0.385	0.388
Adapt_c scheme: $I_{BBN}Q_{crit}$	106	0.642	0.602	0.591	0.587
Speaker1_c scheme: $I_{BBN}Q_{crit}$	119	**0.696**	0.638	0.629	0.624
Speaker2_c scheme: $I_{BBN}Q_{crit}$	151	0.652	0.603	0.596	0.592
Adapt_c scheme: $I_{BBN}Q = 0.5$	106	0.382	0.254	0.215	0.199
Speaker1_c scheme: $I_{BBN}Q = 0.5$	119	**0.509**	0.351	0.315	0.298
Speaker2_c scheme: $I_{BBN}Q = 0.5$	151	0.486	0.347	0.318	0.304

Table 19: Segment based evaluation of clustering & labelling schemes for dominant speaker and dominant environment labelled BNdev96 data partitioned into 553 non-homogeneous segments.

speakers as opposed to 77 true speakers, which denotes that automatic segmentation has allowed 7 speakers to be completely dominated by the others. Furthermore, 6 segments are produced with no foreground speaker, each of which is assigned a new *SIL* speaker.

The clustering schemes detailed in section 4.3.2 are used for assigning speaker turn ID's. It has been concluded from the evaluations of 4.3.3 that the 'speaker1_c' clustering scheme performs the best speaker based clustering. This scheme is used to cluster the automatic segments. The segments in each of the 119 resulting clusters are identified by a unique speaker ID SPK001 through SPK119 respectively.

Secondly, assuming that the approximate number of important speakers to be expected is known to be close to 77, the 'speaker1_c' scheme is minimally adjusted to reduce the number of clusters. The 'speaker1_a_c' scheme increases the allowed re-combinations, while tree growth is held constant. The 'speaker1_b_c' scheme marginally increases the threshold gain required to split in order to grow a smaller tree, while the recombination

	No.Clust	$Norm(I_{RAND})$	$I_{BBN}Q_{crit}$	$I_{BBN}Q = 0.5$
speaker1_c	119	0.396	0.696	0.509
speaker1_a_c	114	0.399	0.699	0.505
speaker1_b_c	91	0.418	0.706	0.479

Table 20: Segment based evaluation of adjusted speaker1_c clustering for dominant speaker, **spkr** labelled BNdev96 data partitioned into 553 non-homogeneous segments.

threshold is held constant. It is observed that the 'speaker1_a_c' produces 114 clusters and 'speaker1_b_c' produces 91 clusters. The clusters from 'speaker1_b_c' scheme is selected and the speaker turns are re-assigned with ID's SPK_B_001 through SPK_B_091.

To assess the success of the initial clustering process, the automatic clusters are evaluated at the segment level according to the dominant labels. For each of the 4 labelling schemes the segment level evaluation results are presented in Table 19. The 'adapt_c' and 'speaker1_c' schemes have produced a larger number of clusters than for the manual segments. The auto segment clustering has produced 119 clusters for 70 dominant speakers, as opposed to 92 clusters for 77 speakers for manual segments. The [0,1] scaled clustering efficiency has decreased slightly from 0.707 to 0.696. However, [-1,1] scaled efficiency shows an increase from 0.44 to 0.51. This may be due to the lowered possibility of producing mixed clusters from having 6 fewer speaker labels. Considering all 3 metrics it can be concluded that the 'speaker1_c' scheme has indeed performed a strong speaker based clustering on the automatic segments. However, there is an environmental influence shown on the clusters as well.

The frame based evaluation of the clustering in Table 21 shows readings very close to the segment based evaluation. The frame based rankings also closely support the dominant-labelled segment based rankings.

The adjusted schemes are evaluated in Table 20. Both Rand and BBN [0,1] scaled efficiency measures show the best scores for the selected 'speaker1_b_c' scheme. However, the [-1,1] scaled BBN measure shows a slight decrease. The selected automatic scheme generates 91 clusters comparable to the 92 clusters generated for the manual segments. The BBN [0,1] scaled efficiency of 0.706 is nearly that of the manually partitioned segment clustering (0.707) for 'speaker1_c'. The BBN [-1,1] scaled measure reading 0.479 shows an improvement over the manual clustering reading (0.436). It is an indication that the number of clusters must not be reduced too much and that a clustering that strikes an appropriate balance between the two BBN metrics must be chosen.

The frame-level evaluation of the adjusted schemes (Table 22) reveal that a speaker turn assignment of a high efficiency of 0.713, compared to the homogeneous speaker clustering efficiency of 0.760, has been achieved. The speaker-environment efficiency readings show that the environment exerts an influence on the clusters and speaker turn assignment may

	No.Clusters	**spkr**	**spkr-env2**	**spkr-env3**	**spkr-music**
No of Labels		71	93	101	105
Adapt_c scheme: $Norm(I_{RAND})$	106	0.252	0.272	0.282	0.282
Speaker1_c scheme: $Norm(I_{RAND})$	119	0.392	0.383	**0.395**	0.394
Speaker2_c scheme: $Norm(I_{RAND})$	151	0.371	0.383	0.381	0.380
Adapt_c scheme: $I_{BBN}\ Q_{crit}$	106	0.672	0.605	0.606	0.609
Speaker1_c scheme: $\eta_{BBN}Q_{crit}$	119	**0.685**	0.596	0.599	0.599
Speaker2_c scheme: $\eta_{BBN}Q_{crit}$	151	0.637	0.559	0.564	0.564
Adapt_c scheme: $\eta_{BBN}Q = 0.5$	106	0.442	0.285	0.271	0.267
Speaker1_c scheme: $\eta_{BBN}Q = 0.5$	119	**0.490**	0.292	0.281	0.274
Speaker2_c scheme: $\eta_{BBN}Q = 0.5$	151	0.458	0.281	0.381	0.270

Table 21: Frame based evaluation of clustering & labelling schemes for BNdev96 data partitioned into 553 non-homogeneous segments.

	No.Clust	$Norm(I_{RAND})$	$I_{BBN}Q_{crit}$	$I_{BBN}Q = 0.5$
speaker1_c	119	0.392	0.685	0.490
speaker1_a_c	114	0.395	0.689	0.487
speaker1_b_c	91	0.407	0.713	0.490

Table 22: Frame based evaluation of adjusted speaker1_c clustering for **spkr** labelled BNdev96 data partitioned into 553 non-homogeneous segments.

be improved by using speech de-noising prior to clustering.

5 Confidence Levels

The motivation for the computation of confidence measures is to be able to detect possible errors in the output of a speech recognition system. Using confidence measures, individual words can be labelled as either correct or false. This additional information about the recognition output can be used for many applications and will be used in the framework of maximum-likelihood-linear regression and unsupervised training of acoustic models within the Coretex project.

With the rising number of different application areas for speech recognition technology, the demand for the ability to spot erroneous words also increases. In this context confidence measures can be used to label individual words in the output of the speech recognition system with either *correct* or *incorrect* thus enabling the system and subsequent modules to spot the position of possible errors in the output automatically.

This additional assessment of the word sequence produced by the speech recognition system has been and can be used in a variety of different applications:

- Confidence measures can be applied to unsupervised training and adaptation algorithms, e.g. vocal tract length normalization, maximum likelihood linear regression [21], and training of acoustic models on automatically generated transcriptions [32]. In all of these cases confidence measures can be used to confine the algorithms to those speech segments whose transcription is most probably correct.
- Another application is the decoding of the speech signal itself. In [31, 33], the authors use confidence measures directly to improve the performance of the speech recognition system.

In the following we will try to motivate our work by discussing why the computation of confidence measures in a speech recognition system is in fact a problem. The fundamental rule in all statistical speech recognition systems is Bayes' decision rule which is based on the posterior probability $p(w_1^M|x_1^T)$ of a word sequence $w_1^M = w_1, \ldots, w_M$, given a sequence of acoustic observations $x_1^T = x_1, \ldots, x_T$. That word sequence $\left\{w_1^M\right\}_{opt}$ which maximizes this posterior probability also minimizes the probability of an error in the recognized sentence:

$$\begin{aligned}\left\{w_1^M\right\}_{opt} &= \arg\max_{w_1^M} p(w_1^M|x_1^T) && (23)\\ &= \arg\max_{w_1^M} \frac{p(x_1^T|w_1^M)\cdot p(w_1^M)}{p(x_1^T)} && (24)\\ &= \arg\max_{w_1^M} p(x_1^T|w_1^M)\cdot p(w_1^M), && (25)\end{aligned}$$

where $p(w_1^M)$ denotes the language model probability, $p(x_1^T|w_1^M)$ the acoustic model probability and $p(x_1^T)$ the probability of the acoustic observations. Strictly speaking, the maximization is also over all sentence lengths M.

If these posterior probabilities were known, the posterior probability $p(w_m|x_1^T)$ for a specific word w_m could easily be estimated by summing up the posterior probabilities of all sentences w_1^M containing this word at position m. This posterior word probability could directly be used as a measure of confidence.

Unfortunately, the probability of the sequence of acoustic observations $p(x_1^T)$ is normally omitted since it is invariant to the choice of a particular sequence of words. The decisions during the decoding phase are thus based on unnormalized scores. These scores can be used for a comparison of competing sequences of words, but not for an assessment of the probability that a recognized word is correct. This fact, and in other words the estimation of the probability of the acoustic observations, is the main problem for the computation of confidence measures.

The posterior probability for a word hypothesis can be computed on the basis of word graphs. In the style of the forward-backward algorithm we compute the forward probability and the backward probability for a word hypothesis and combine both probabilities into the posterior probability of this hypothesis. In contrast to the forward-backward algorithm on a Hidden-Markov-Model state level, the forward-backward algorithm is now based on a word hypothesis level.

These posterior hypothesis probabilities turned out to perform poorly as a confidence measure. In fact, this observation is not surprising since the fixed starting and ending time of a word hypothesis determine which paths in the word graph are considered during the computation of the forward-backward probabilities. Usually, several hypotheses with slightly different starting and ending times represent the same word and the probability mass of the word is split among them. In order to solve this problem, the posterior probabilities of all those hypotheses which represent the same word have to be summed up. More details are given in [34].

Experimental evidence clearly shows that posterior word probabilities outperform alternative confidence measures, i.e. the acoustic stability and the hypothesis density. Additional experiments prove that the estimation of posterior word probabilities on word graphs yields better results than their estimation on N-best lists. The relative reduction in confidence error rate ranges between 19% and 35% on different corpora (ARISE, NAB 20k and 64k, Verbmobil and Hub4 Broadcast News) using a trigram language model and the best posterior probability based confidence measure. The relative reduction was highest for corpora which are commonly regarded as difficult, consisting of spontaneous speech. For these corpora, the advantage of the confidence measures based on word graph posterior probabilities was also highest compared to the other confidence measures. It is interesting to note that this improvement is achieved with a single confidence measure and not with a vector of numerous features which can be extracted from a word graph.

In the framework of Coretex, confidence measures are used for unsupervised training (see Deliverable D1.3, “Report on light supervision techniques) and for improved speaker adaptation (see Deliverable D1.1, “Report on Genericity and Adaptability”).

6 Conclusions

In this report a comparative study of techniques for Acoustic Meta-Data mark up is presented. The work is presented under four sections.

Segmentation

As a pre-segmentation task the acoustic data stream is segmented and marked as speech, non-speech music, or non-speech noise sections, in order to segregate speech from non-speech data. The non-speech sections are discarded. Of the Maximum Likelihood and Maximum Mutual Information based decoders built for this classification task, the MMI based decoder achieves superior performance giving only 0.89% speech loss of German Broadcast News data. In addition 98.84% pure clean speech segmentation is achieved.

Secondly, segmentation based on several model selection criteria which perform transition detection is investigated for Italian Broadcast News data. The aim of the experiments is to evaluate several model selection schemes that detect spectral changes that occur within the signal, which are due to channel and source switches. The segments are produced at the marked points of spectral change in the signal, to produce segments that are homogeneous with respect to the foreground speech and background noise sources. It is demonstrated that several of the techniques evaluated - particularly those based on BIC, CAIC and MDL criteria - can be tuned to achieve segmentation of high (about 98-99%) precision.

This level of mark-up of non-speech sections and homogeneous segments with respect to speech and background environment in the acoustic data stream is useful for determining and improving speech recognition accuracy.

Speaker and Environment Labelling

Further mark-up of the cleaned speech segments is performed by recognition of the foreground speaker and background noise conditions. This meta data is useful for speaker and environmental adaptation and, in turn, the improvement of recognition accuracy of speech recognition systems. Other applications exist such as voice recognition security systems.

Speaker Identification on Italian Broadcast News is performed using 3 techniques. Firstly, GMM classifiers perform Maximum Likelihood classification assuming perfect segmentation. Secondly, the *loop* technique performs a Viterbi-based classification algorithm that allows recovering boundaries in segments not previously detected, thereby allowing recovery from segmentation errors. Thirdly, a *hierarchy* technique combines the above by first performing *loop*-classification into generic audio classes, then performing ML classification into speakers. The experiments show that the *hierarchy* technique is able to achieve a maximum 81.9% frame classification accuracy. In IBNC 69 of the 73 test speakers are seen in the training data; the unseen speakers are recognized as an 'unknown' class.

Speaker Identification for the US Broadcast News corpus is somewhat different because only 2 of the 77 test speakers, who are dominant speakers of only 84 of the 553 automatic test segments, are seen in the training data. The system is built based on a ML trained UBM model MAP adapted for the top 50 speakers. The system correctly recognizes 70 (83.3%) of the seen speaker segments and correctly marks 457 (97.4%) of the 'unknown' speaker segments. The 'unknown' speaker segments are clustered to achieve a 0.659 segment level (0.692 frame level) speaker clustering efficiency.

Maximum Likelihood GMM classifiers are built to assign background noise conditions to the unlabelled automatic US-BN data segments with speech in the foreground. The classifier achieves a noisy/clear environment assignment of 90.5% frame level accuracy (88.6% dominant environment labelled segment level accuracy) for American English Broadcast News data. It is less successful with further discrimination between noise loudness levels or music in the background.

The recognition results on US-BN data reflect errors in segmentation as well as recognition since no segmentation error recovery techniques are employed.

Cluster Assessment for Speaker-Environment Turns

It is only possible to correctly recognize speakers that are seen in the training data. For 'unknown' speakers previously unseen, clustering techniques are used to group together and mark segments of speech by the same speaker. The cluster assessment framework is extended to assess the success of grouping according to speakers *and* noise environments. Two metrics are adapted for the joint evaluation of clustering and labelling schemes and the merits of the metrics are discussed.

The investigations reveal that, while both metrics are suitable for the evaluation of clustering schemes, the BBN efficiency metric is preferred for the evaluation of labelling schemes. For homogeneous segments at frame level, the clustering scheme that grows a moderate size tree with re-combinations achieves a speaker clustering of a high efficiency of 0.760, and a speaker-environment clustering of a closely high efficiency of 0.715 through a noisy/clear environment. This result shows that the environment exerts a high level of influence on speech data clustering. For the automatic US-BN data segments, the application of the same clustering scheme with minimal adjustments achieves a speaker turn assignment of a comparatively high efficiency of 0.713 at frame level. The results suggest that clustering for speaker turns can be improved by prior speech de-noising.

Confidence Levels

Confidence level mark up of a speech recognition result provides an assessment of recognition accuracy and, in turn, allows further improvement of speech recognition accuracy.

The posterior probability of *word hypotheses*, estimated in a manner similar to the estimation of the posterior probabilities of HMM hidden states, are summed up for the same word to achieve *posterior word probabilities*. Posterior word probabilities are shown to outperform other confidence measures, and reduce errors in confidence estimations by 19%-35% for varied corpora.

A number of further research directions suggested by the results of this work include possible improvement of background noise condition mark up by using MMI based classifiers and segmentation error-recovery techniques, investigation of other metrics appropriate for cluster assessment, and investigation of speech recognition improvements achievable by the combined use of generated acoustic meta data.

7 References

[1] H. Akaike: On entropy maximization principle. In P. R. Krishnaiah, editor, *Applications of Statistics*, pages 27–41. North-Holland, Amsterdam, Nederlands, 1977.

[2] R.A. Baxter: *Minimum Message Lenght Inference: Theory and Applications*. PhD thesis, Department of Computer Science Monash University, Clayton, Victoria, Australia, 1996.

[3] H. Bozdogan: Model selection and the Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[4] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani: A system for the segmentation and transcription of Italian radio news. *Proceedings of RIAO Content-Based Multimedia Information Access*, Paris, France, 2000.

[5] M. Cettolo: Segmentation, classification and clustering of an Italian broadcast news corpus. *Proceedings of the RIAO International Conference*, Paris, France, 2000.

[6] M. Cettolo and M. Federico: Model selection criteria for acoustic segmentation. *Proc. of the ISCA Automatic Speech Recognition Workshop*, Paris, France, 2000.

[7] M. Cettolo: Speaker Tracking in a Broadcast News Corpus. *Proc. of the 2001: A Speaker Odyssey - The Speaker Recognition (ISCA) Workshop*, Chania (Crete), Greece, 2001.

[8] S.S. Chen and P.S. Gopalakrishnan: Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. *Proc. of the DARPA Broadcast News Transcr. & Understanding Workshop*, Lansdowne, VA, 1998.

[9] J.H. Conway, N.J.A. Sloane: *Sphere Packing, Lattices and Groups*. Springer Verlag, Berlin, Germany, 1988.

[10] R.O. Duda, P.E. Heart D.G. Stork: *Pattern Classification* 2nd Edition 2001, John Wiley & Sons

[11] M. Federico, D. Giordani, P. Coletti: Development and Evaluation of an Italian Broadcast News Corpus. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000.

[12] J.-L. Gauvain, L. Lamel, G. Adda, and M. Jardino: The LIMSI 1998 Hub-4E transcription system. *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA, 1999.

[13] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, and S.J. Young: Segment generation and clustering in the HTK broadcast news transcription system. *Proc. of the DARPA Broadcast News Transcr. & Und. Workshop*, Lansdowne, VA, 1998.

[14] L. Hubert, P. Arabie: Comparing Partitions. *Journal of Classification*, Vol.2 pp.193-218, 1985

[15] S.E. Johnson: Who Spoke When? - Automatic Segmentation and Clustering for Determining Speaker Turns. *Proc EuroSpeech*, Volume 5 pp.2211-2214, 1999

[16] S.E. Johnson, P.C. Woodland: Speaker Clustering Using Direct Maximization of the MLLR-Adapted Likelihood. *Proc. ICSLP*, Vol.5 pp.1775-1779, 1998

[17] R. Kuhn, P. Nguyen , J.-C. Junqua, L. Goldwassar, N. Neidzeilski, S. Fincke, K. Field, M. Contolini: Eigenvoices for Speaker Adaptation. *Proc. ICSLP* pp.1771-1774, 1998

[18] D. Liu and F. Kubala: Fast speaker change detection for broadcast news transcription and indexing. *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 1031–1034, Budapest, Hungary, 1999.

[19] W. Macherey, H. Ney: Towards Automatic Corpus Preparation for a German Broadcast News Transcription System *Proc. ICASSP*, Orlando, Florida, May 2002

[20] A. Martin, G. Doddington, T. Kamm, M.Ordowski & M. Przybocki (1997): The DET Curve in Assessement of Detection Task Performance. *Proc. Eurospeech 1997*, Vol. 4, pp 1895 - 1897

[21] M. Pitz, F. Wessel, H. Ney: Improved MLLR Speaker Adaptation Using Confidence Measures for Conversational Speech Recognition, *Proc. 6th ICSLP*, Beijing, China, pp. 157-159, October 2000.

[22] D.A. Reynolds and R.C. Rose: Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3(1):72–83, 1995.

[23] D.A. Reynolds: Comparison of Background Normalization Methods for Text-Indpendent Speaker Verification. *IEEE Trans. Speech and Audio Processing*, 3(1):72–83, 1995.

[24] J. Rissanen: Stochastic complexity. *Journal of the Royal Statistical Society*, 49(3):223–239, 1987.

[25] G. Schwarz: Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[26] G.A.F. Seber: *Multivariate Observations*. John Wiley & Sons, New York, NY, 1984.

[27] A. Solomonoff, A. Meilke, M. Schmidt, H. Gish: Clustering Speakers by their Voices. *Proc. ICASSP*, Vol.2 pp.757-760, 1998.

[28] M.S. Srivastava and E. M. Carter: *An Introduction to Applied Multivariate Statistics*. North-Holland, New York, NY, 1988.

[29] C.S. Wallace and P.R. Freeman: Estimation and inference by compact coding. *Journal of the Royal Statistical Society, B*, 49(3):240–265, 1987.

[30] S. Wegmann, P. Zhan, and L. Gillick: Progress in broadcast news transcription at Dragon Systems. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 33–36, Phoenix, AZ, 1999.

[31] F. Wessel, R. Schlüter, H. Ney: Using Posterior Probabilities for Improved Speech Recognition, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2000*, Istanbul, Turkey, pp. 1587-1590, June 2000

[32] F. Wessel, H. Ney: Unsupervised training of acoustic models for large vocabulary continuous speech recognition *Proc. Automatic Speech Recognition Workshop 2001*, Madonna di Campiglio, Trento, Italy, December 2001.

[33] F. Wessel, R. Schlüter, H. Ney: Explicit word error minimiza-tion using word posterior probabilities *Proc. International Conference on Acoustics, Speech, and Signal Processing 2001*, Salt Lake City, UT, USA, vol. 1, pp. 33-36, May 2001.

[34] F. Wessel, R. Schlüter, K. Macherey, H. Ney: Confidence measures for large vocabulary continuous speech recognition, *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288 298, March 2001.

[35] J.T. Wickramaratna, M.J.F. Gales, P.C. Woodland: Assigning Speaker Turns and Noise Conditions in Broadcast News. *Tech. Report CUED/F-INFENG/TR.425*, Cambridge University, Cambridge, UK

[36] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Neisler, A. Tuerk, E.W.D. Whittaker, S.J. Young: The 1997 HTK Broadcast News Transcription System *Proc. DARPA Broadcast News Transcription and Understanding Workshop 1998*, pp. 41-48, Lansdowne, Virginia