Digital Signal Processing **10**, 113–132 (2000)
doi:10.1006/dspr.1999.0357, available online at http://www.idealibrary.com on IDEAL®

Local Normalization and Delayed Decision Making in Speaker Detection and Tracking

Johan Koolwaaij and Lou Boves

Automatic Acoustic Recognition Technologies, University of Nijmegen, Nijmegen 6500HD, The Netherlands

E-mail: koolwaaij@let.kun.nl; boves@let.kun.nl

Koolwaaij, Johan, and Boves, Lou, Local Normalization and Delayed Decision Making in Speaker Detection and Tracking, *Digital Signal Processing*, **10** (2000), 113–132.

This paper describes A2RT's speaker detection and tracking system and its performance on the 1999 NIST speaker recognition evaluation data. The system does not consist of concatenated modules such as, for instance, silence–speech detection, handset and gender detection, and finally speaker detection or tracking, where each module builds on the hard decisions from previous modules, but rather applies the principle of delayed decision making and postpones all hard decisions until the final stage of the detection process. This paper focuses on two important locality issues in detecting or tracking speakers in a telephone conversation, for which the speaker change frequency is usually high. First, channel estimation needs sufficiently long but homogeneous segments. Several kinds of local channel normalization are compared in this paper. Second, local estimation of speaker likelihoods critically depends on the segmentation of the conversation. Our experiments show that a global level of segmentation really improves speaker tracking performance, whereas a more detailed segmentation is needed for speaker detection, because likelihood computation over clusters of segments depends on the purity of the segments. Furthermore, choosing the appropriate type of channel normalization can give a small but consistent improvement in speaker tracking performance.

Key Words: decision making; segmentation; speaker detection; speaker tracking

0. INTRODUCTION

Speaker detection and speaker tracking are new tasks in the framework of the NIST speaker recognition evaluation campaign. However, similar tasks have been approached [2] in the context of air traffic control and, in parallel, in the

1051-2004/00 $35.00

Broadcast News task [1, 3]. The context created by the NIST campaign seems to dictate new approaches, adapted to the specific requirements of the task. Thus far, the performance of the two-speaker detection and speaker tracking tasks has been substantially lower than that of the conventional one-speaker detection task. In the latter task, equivalent to speaker verification, one can use the a priori information that all speech comes from exactly one speaker [8]. One of the advantages of one-speaker detection is that all normalization parameters (such as the cepstral mean for channel normalization) and all likelihood values can be computed over the whole utterance; for two-speaker detection and speaker tracking the conversation must be segmented into homogeneous segments before normalization parameters, likelihood values, etc., can be computed.

In the absence of a priori information on speaker segmentation one faces two problems: (i) segmentation errors affect the computation of the parameters of the channel normalization as well as the likelihood normalization; (ii) systems based on a concatenation of modules (such as silence–speech detection, handset and gender detection, and finally speaker detection) suffer from the error propagation effect: hard decisions in the segmentation process cannot be corrected later in the speaker detection process. For example, segmenting a conversation into female and male segments and then applying speaker detection on the segments assigned to the gender of the claimed speaker only may lead to a high speaker miss rate due to errors in the male–female segmentation. Moreover, concatenated systems are hard to tune, because a threshold setting in one module may interact with a threshold setting in another.

In this paper solutions are proposed and tested for both problems. First, different procedures for channel normalization will be compared. These procedures allow for different trade-offs between short segments, which are very likely related to a single channel but may not allow reliable estimation of the relevant parameters, and longer segments, which suffer less from the lack of data problem, but possibly at the cost of joining contributions from different channels into a single segment. Second, we propose a segmentation algorithm that postpones its decision until the moment at which no additional information can become available. The segmentation algorithm does not use any explicit rule, like minimum segment duration; such rules may help on Broadcast News data [3], but on the highly interactive Switchboard conversations different rules or different rule settings of the parameters in the rules will be needed.

1. THE TASKS

The 1999 speaker recognition evaluation is part of an ongoing series of annual evaluations conducted by NIST. All evaluation data come from Switchboard corpora containing telephone conversations by English-speaking subjects. Testing data include 723 Switchboard conversations created by summing the separately recorded telephone channels and 1972 monologues created by removing silence from one of the telephone channels.

This paper reports on three different speaker recognition tasks. *One speaker detection* tests the hypothesis that a target speaker is speaking given a speech utterance containing only one speaker. For each of the 1972 monologues there are 11 hypothesized speakers. One is the actual speaker; the other 10 are imposters, all of the same sex. The total amount of speech being tested is about 186 h in $1972 \times 11 = 21,692$ tests. *Two-speaker detection* tests the hypothesis that a target speaker is speaking given a conversation between two speakers. For each of the 723 conversations there are 22 hypothesized speakers. One or two are actual speakers; the others are imposters; half of the speakers are male and half are female. This makes a total of 15,906 tests and about 264 h of conversation to be tested. *Speaker tracking* tests the hypothesis that a target speaker is speaking at time t given a conversation between two speakers. The test set is exactly the same as for two-speaker detection.

2. A2RT'S SPEAKER DETECTION AND TRACKING SYSTEM

2.1. Features

Parameterization is based on 25.6-ms Hamming windows, with a 10-ms window shift. For each frame 12 LPC cepstra (with 16th order LPC) and log energy are computed; the feature vector is formed by appending the deltas and delta–deltas of the 13 coefficients, making for a total of 39 features. The preemphasis coefficient is set to 0.97. The delta coefficients are computed over a 50-ms window, and the delta–delta coefficients over a 90-ms window.

Because a Switchboard conversation is the sum of the two separately recorded channels with possibly different channel characteristics, cepstral mean should not be calculated over the whole conversation. Instead, we subtract the cepstral mean $\bar{\mu}_c$ and divide by the cepstral standard deviation $\bar{\sigma}_c$, where $\bar{\mu}_c$ and $\bar{\sigma}_c$ are calculated over a limited time window [10]. The time window should be long enough to obtain reliable estimates of the cepstral mean and standard deviation, yet contain the speech of only one speaker. It would be optimal to first separate the channels then normalize them; however, at this moment there is no blind channel separation algorithm available. In this paper we use three different procedures with different time window lengths. The first procedure segments the signal on the basis of overall energy; therefore, it uses variable duration windows. The two remaining procedures use sliding windows of 250 and 1000 ms, respectively, corresponding to the 5th and 50th percentile of the segment duration distribution in the one-speaker data. A 250-ms window is likely to be homogeneous but short, while a 1000-ms window contains enough data to allow for reliable estimates of the channel normalization (CN) parameters, provided that all data come from a single channel, which may not be true.

CN1: The data are segmented in segments with a minimum duration of 1000 ms using an energy threshold to detect silence–speech transitions. All cepstral vectors in a segment are normalized by $\bar{\mu}_c$ and $\bar{\sigma}_c$ computed over that

segment. (Note that this segmentation is based on energy only and does not use any silence, speech, or speaker model.)

CN2: All cepstral vectors are normalized by $\bar{\mu}_c$ and $\bar{\sigma}_c$ computed for a sliding rectangular window with a width of 1000 ms.

CN3: All cepstral vectors are normalized by $\bar{\mu}_c$ and $\bar{\sigma}_c$ computed for a sliding rectangular window with a width of 250 ms.

Channel normalization is then performed by rescaling the cepstral vector $\bar{c}$ to the standard normal domain:

$$\mathrm{CN}(\bar{c}) = \frac{\bar{c} - \bar{\mu}_c}{\bar{\sigma}_c}.$$

Channel normalization is applied to both the training and the testing data to ensure that there is no unpredictable mismatch between the speaker model and the speaker's testing data.

2.2. Modeling

Anti-speaker modeling. To normalize the target speaker likelihood a gender- and handset-dependent world model is trained. Possible gender categories are **m**ale and **f**emale, and possible handset categories are **carb**on button and **elec**tret. First, a set of initial model parameters is computed using a segmental K-means procedure. Then, the parameters are further reestimated using the Baum–Welch training algorithm. All models are GMMs with 128 mixtures, which gave optimal speaker verification performance when speaker models are trained on 2 min of speech. The world models, together with a silence model, are trained using a total of about 2 h of speech data from 306 different speakers taken from the 1998 evaluation data.

Speaker modeling. There are 309 female and 230 male speakers in the database. Each target speaker is modeled by one text-independent GMM with the same topology as the world models. To enroll each speaker two conversations from the same telephone number are available; between 55 and 65 s of speech is taken from each conversation, making for a total of about 2 min of training data per speaker. All training speech was used to train the target model by a Baum–Welch reestimation algorithm bootstrapped from the world model corresponding to the gender of the target and the handset type used for the recording of the training utterances.

Variance modeling. The variances of the target model are trained using the target's training data, but a variance floor vector [6] is set, which prevents variances from becoming too small due to overfitting of the models on the speaker's training data. The variance floor vector is gender and handset dependent and is trained on the same data used for training the world and silence models.

Normalization. All training data for the client models are from one handset type, so each speaker model contains not only speaker information but also information about the handset used during training. To normalize for this undesirable bias we use the hnorm technique [9]. During testing

all log likelihood ratio (LLR) scores are transformed such that the imposter score distribution resembles the standard normal distribution, using a mapping that is speaker, gender, and handset dependent. The parameters used for this mapping are computed per target model using 30 min of speech data from 151 different speakers from the 1998 evaluation. The parameter set can be denoted as $\{\mu^T_{H,G}, \sigma^T_{H,G}\}$, where T is the target speaker, H is the handset used by the segment speaker and G is the gender of the segment speaker.

2.3. Segmentation

For the one-speaker detection task one side of a conversation is provided as testing data. The test data set is provided by NIST with silences removed and duration varies from a few seconds to 1 min (on average 31 s). Prior knowledge of the speaker's gender and the handset type is permitted; however, we did not use this information in our experiment. The main reason for not using such information is that we prefer to rely on the decision of our own gender and handset models; in addition, in real applications prior knowledge about gender and handset may not be available. For the two-speaker detection and the speaker tracking tasks the two sides of the conversation are summed. Silence is still present and duration is between 59 and 61 s. Prior information about neither the handset mixture nor the gender mixture is permitted.

Chopping a Switchboard conversation into homogeneous segments is far from trivial. One approach is to use a silence–speech segmentation, which works reasonably well when it is safe to assume that there is an intervening silence between two speakers, as in the Broadcast News task [3]. However, this is too simple an approach for the Switchboard conversations, which contain a significant amount of overlapping speech. Also, background noise may mask interspeaker silences. Our approach is to use all available information sources to perform the data chopping task. Per test utterance six likelihood streams are computed: one stream for the silence model, four streams for possible world models (F-carb, F-elec, M-carb, M-elec), and one stream for the target model. The likelihood streams are computed on a per-frame basis. The segmentation task is performed by a dynamic programming algorithm that uses the following types of detection:

1. Silence–speech detection tests the hypothesis that a segment belongs to one of the four general speech models versus the hypothesis that it is a silence segment.
2. Handset detection tests the hypothesis that the handset used to record a segment was an electret versus the hypothesis that it was a carbon-button handset.
3. Gender detection tests the hypothesis that a segment is produced by a male versus the hypothesis that it is produced by a female speaker.
4. Speaker detection tests the hypothesis that a segment is from a (known) target speaker versus the hypothesis that it is from an (unknown) non-target.

The more the two speakers in a conversation have in common, the smaller the differences between segments pertaining to different speakers and the

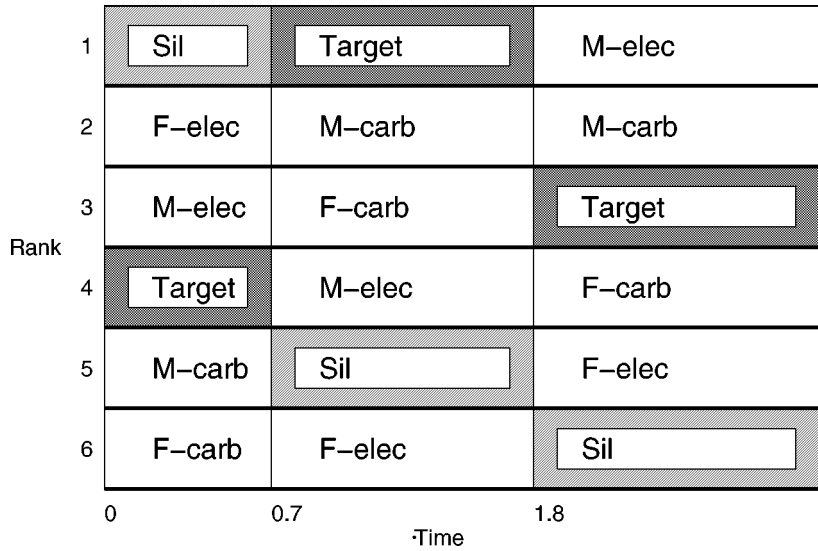

FIG. 1. Example of DP output. On the horizontal axis is time in seconds; on the vertical axis is the rank of the models. The target model is dark gray, the silence model is light gray, and the world models are white.

more difficult the segmentation task. In these cases a more general transition detection algorithm would be helpful. It would be interesting to compare our approach to transition detection by means of the Bayesian information criterion (BIC) [1], which is very successful on Broadcast News data, on the Switchboard conversations with high speaker change frequency.

Dynamic programming. A dynamic programming (DP) algorithm takes care of the segmentation of the conversation. Input are the six likelihood streams described in the previous subsection. The raw likelihood values are very noisy, so smoothing is applied using a Hamming window with a width of 100 ms. The DP applies a maximum likelihood principle together with a penalty for short segments (PSS). Let $\{x_1, x_2, \ldots, x_n\}$ be the set of cepstral feature vectors; the DP algorithm then finds the segmentation S, which maximizes

$$\mathcal{L}(S) = \log P(x_1, x_2, \ldots, x_n|S) + \mathrm{PSS} \cdot L(S),$$

where $L(S)$ is the mean length of the segments in segmentation S. By increasing PSS, long segments are favored and short segments penalized. Therefore, the penalty for short segments controls the level of detail of the segmentation. This penalty can be compared to the word entrance penalty in speech recognition tasks. The output is the optimal segmentation and an N-best list of models per segment. The segment boundaries are kept fixed, but no hard decisions concerning the assignment of segments to a certain model are made yet. The N-best list per segment makes it possible to reconstruct all possible paths through model space, with the restriction of a given set of segment boundaries.

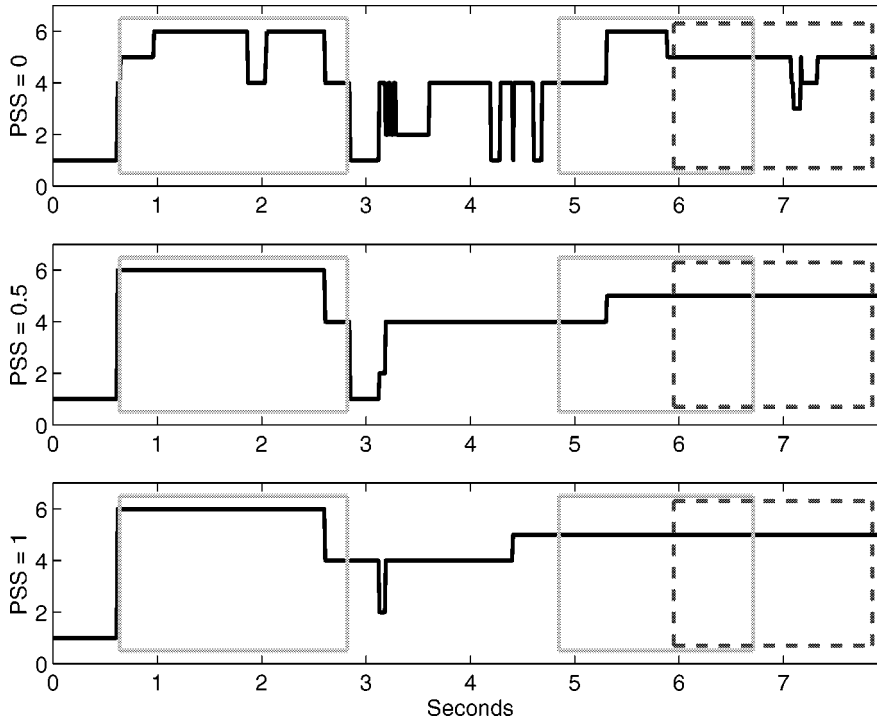

FIG. 2. Example of speaker tracking in a telephone conversation. Both speakers are male using electret handsets. The solid boxes show where the target speaker is speaking according to the NIST answer keys, the dashed box is the non-target speaker. The solid thick line represents the rank 1 solution of the DP output: on the vertical axis 0 is silence; 1–5 are the F-carb, F-elec, M-carb, and M-elec world models, respectively; and 6 is the target. Results are shown for the PSS values 0, 0.5, and 1.

Figure 1 shows an artificial example of the DP output with fixed segment boundaries and full N-best information per segment. This example will be used to clarify the likelihood measures in the next section. Figure 2 is an example of (a part of) a conversation between two males using electret handsets. The example shows that an increasing PSS value gives better detection of the first target segment, while the second target segment, which is partly overlapped by another speaker, is no longer detectable using a more global segmentation. Also, with PSS $= 0.0$ the segments have a higher probability of containing speech from only one speaker (or channel), but they may have very short lengths resulting in potentially unreliable time-normalized likelihood estimates. With higher PSS values the segments have longer duration but they are not always homogeneous (in the sense that they only contain speech from one speaker or channel). The question is what the optimal setting is for one-speaker detection, two-speaker detection, and speaker tracking. Furthermore, is the optimal setting of the parameters of the DP algorithm and the channel normalization the same for all three tasks? Can we understand why or why not?

2.4. Likelihood Measures

This section describes the likelihood measures used for the three different tasks one-speaker detection, two-speaker detection, and speaker tracking. Let $\mathrm{LL}(M,t)$ be the log likelihood score for a given model M at time t.

One-speaker detection. For one-speaker detection we can make use of the prior information that all speech in a test utterance comes from exactly one speaker. Therefore, we use the segmentation only to select the best fitting world model, which probably, but not necessarily, corresponds to the gender of the speaker and to the handset type used while recording the segment. If we define the nonsilence (NS) segments as those segments for which silence does not have rank 1 in the segmentation, then the presence P of a world model W can be defined as the percentage of all NS frames with W having the highest rank of all world models. The best fitting world model is the world model with the highest P value. This world model (World_1) is used to normalize the target speaker likelihood by subtracting the likelihood for World_1 (and thus creating a log likelihood ratio), hnorming the LLR, and integrating over all segments where the target has rank 1 (T). Duration normalization is accomplished by dividing by the total duration of the NS segments.

$$\mathrm{LLR} = \frac{\displaystyle\int_T \mathrm{hnorm}(\mathrm{LL}(\mathrm{Target},t) - \mathrm{LL}(\mathrm{World}_1,t))\,dt}{\displaystyle\int_{\mathrm{NS}} dt}. \tag{1}$$

Note that the hnorm is speaker, gender, and handset dependent: the appropriate hnorm parameters are the μ and σ for this model speaker (Target) using the handset and gender corresponding to the world model (World_1). Note that LLR can also be written as

$$\mathrm{LLR} = \frac{\displaystyle\int_T \mathrm{hnorm}(\mathrm{LL}(\mathrm{Target},t) - \mathrm{LL}(\mathrm{World}_1,t))\,dt}{\displaystyle\int_T dt} \cdot \frac{\displaystyle\int_T dt}{\displaystyle\int_{\mathrm{NS}} dt}, \tag{2}$$

in other words as a product of a mean log likelihood ratio over T and a duty cycle (the percentage of time the target is detected in NS). Both factors contain speaker presence information; in [4] it was shown that multiplication of these factors yields best performance. In the example in Fig. 1, NS corresponds to the interval $(0.7, 3)$ and M-carb is the world model with the highest presence (P is slightly more than 50%).

Two-speaker detection. For two-speaker detection we select the two world models with the highest presence in the segmentation (World_1 and World_2). We use the information that two speakers are engaged in conversation and first compute the log likelihood ratio over $T \cap W_1$, where W_1 is defined as those speech segments for which World_1 has a higher rank than World_2 in the segmentation and T are those segments with the target model having rank 1. Second, we compute the log likelihood ratio over $T \cap W_2$ with W_2 defined as those

speech segments for which World_2 has a higher rank than World_1. The final log likelihood ratio is the maximum of these two ratios:

$$\text{LLR}_1 = \frac{\int_{T \cap W_1} \text{hnorm}(\text{LL}(\text{Target}, t) - \text{LL}(\text{World}_1, t))\, dt}{\int_{W_1} dt} \tag{3}$$

$$\text{LLR}_2 = \frac{\int_{T \cap W_2} \text{hnorm}(\text{LL}(\text{Target}, t) - \text{LL}(\text{World}_2, t))\, dt}{\int_{W_2} dt} \tag{4}$$

$$\text{LLR} = \max\{\text{LLR}_1, \text{LLR}_2\}. \tag{5}$$

Equations (3)–(5) assume that all conversations are between speakers belonging to two different gender–handset combinations. This is obviously wrong when two speakers of the same gender using the same handset type are conversing. Then, the set W_1 becomes large with respect to the set W_2. In theory, LLR_2 can become an unstable measure when the set W_2 contains very few frames and hence the integration interval becomes too small to obtain reliable estimates. In practice, however, this is not a significant problem, because the total duration of the W_2 segments almost never drops below a critical value of 500 ms using conversations with a duration of 1 min. In the example in Fig. 1, World_1 is M-carb and World_2 is M-elec. It follows that $W_1 = (0.7, 1.8)$ and $W_2 = (1.8, 3.0)$, and so in this case $T \cap W_1 = (0.7, 1.8)$ and $T \cap W_2 = \emptyset$.

Speaker tracking. For speaker tracking all segments with the target model as rank 1 option are marked as target, and all other segments as nontarget. The log likelihood ratio is computed per speech segment S,

$$\text{LLR} = \frac{\int_S \text{hnorm}(\text{LL}(\text{Target}, t) - \text{LL}(\text{World}_1^S, t))\, dt}{\int_S dt}, \tag{6}$$

where World_1^S is the world model with the highest presence for the segment S. Silence segments get a zero score, which is the mean score for imposters. All segments with the target having rank 1 are marked as target, so in Fig. 1 only the interval $(0.7, 1.8)$ would be marked as target.

2.5. Evaluation

All three tasks were evaluated using the same detection cost function (DCF), which is a linear combination of the miss rate and the false alarm rate,

$$\begin{aligned}\text{DCF} = {} & C_{\text{Miss}} \times P(\text{Miss}|\text{Target}) \times P(\text{Target}) \\ & + C_{\text{FalseAlarm}} \times P(\text{FalseAlarm}|\text{Non-Target}) \times P(\text{Non-Target}),\end{aligned} \tag{7}$$

where C is the cost of a detection error ($C_{\text{Miss}} = 10$ and $C_{\text{FalseAlarm}} = 1$) and $P(\text{Target})$ is the prior probability for a target, equal to 1%. For more detailed information regarding the detection cost function see [5].

TABLE 1

Detection Performance in Percentage Error for Gender and Handset over All 1972 Utterances for the One-Speaker Detection Task

CN type	Handset only	Gender only	Handset plus gender
1	6.5%	2.3%	7.1%
2	8.3%	2.1%	9.3%
3	7.1%	2.2%	7.9%

3. DETECTION AND TRACKING RESULTS

3.1. Gender and Handset Detection

Although the A2RT system does not apply explicit gender and handset detection to the one-speaker utterances, we are still interested in the performance on these tasks. Gender detection is done by determining the presence of the male world models (P_m) and the presence of the female world models (P_f) from the DP output. (For definition of P see Section 2.4.) If $P_m > P_f$, we are dealing with a male utterance; otherwise we are dealing with a female utterance. An analogous procedure is followed for handset detection. All 1972 speech utterances from the one-speaker detection are used to evaluate the performance of the gender and handset detection.

The results are shown in Table 1. The gender detection error rate is close to 2% for all CN types, while handset detection yields error rates ranging from 6.5 to 8.3%, depending on the CN type. The CN dependence for handset detection performance can be explained because channel normalization removes some of the handset characteristics. It should be noted that the "correct" handset labels provided with the NIST 1999 data were obtained using the MIT Lincoln Lab's handset type labeler software [7]. So the handset detection "error" is in fact a difference measure with the handset labels provided by MIT.

3.2. Gender and Handset Mixture Detection

Gender and handset mixture detection in telephone conversations between two speakers is a more complicated task. For gender mixture detection the parameters P_m and P_f are determined for the conversation and compared to an a priori threshold θ:

$$\text{Gender mixture} = \begin{cases} \text{MM} & \text{if } P_m - P_f > \theta \\ \text{MF} & \text{if } -\theta \le P_m - P_f \le \theta \\ \text{FF} & \text{if } P_m - P_f < -\theta. \end{cases}$$

The higher θ, the more mixed gender conversations are detected. The fact that a speaker may be talking from close to 0 s in some conversations to almost 1 min in other conversations implies that a robust setting of the threshold θ is very difficult. A similar procedure is followed for handset mixture detection. Mixture detection results are given in Table 2. The need to estimate additional thresholds makes the tasks of handset and gender mixture detection more error

TABLE 2

Mixture Detection Performance in Percentage Error for Gender and Handset over All 723 Conversations for the Two-Speaker Detection Task

CN type	Handset mix only	Gender mix only	Handset plus gender mix
1	15.6%	10.2%	24.0%
2	14.5%	9.7%	24.9%
3	15.4%	9.4%	26.0%

prone than the gender and handset detection in the one-speaker tasks. The differences in performance between the different CN types are negligible.

3.3. One-Speaker Detection

Mismatch between training and testing conditions is *the* factor that determines the performance in one-speaker detection [9]. Channel mismatch is probably the single most important form of mismatch. Therefore, we present results for different channel (mis)match conditions in the test data:

ALL: all 21,692 tests;

SS: a subset of 10,318 tests under the restriction that the speaker uses the same phone number and the same handset type as during the training recordings;

SD: a subset of 5082 tests under the restriction that the speaker uses a different phone number but the same handset type as during the training recordings;

DD: a subset of 5643 tests under the restriction that the speaker uses a different phone number and a different handset type than during the training recordings.

The channel mismatch increases going from the SS via SD to the DD condition and detection performance drops significantly as is shown in Table 3. The soft-decision DCF values are displayed for the different types of channel normalization.

One-speaker detection experiments with cepstral normalization over the whole utterance were performed as a reference for the CN types 1–3. CN0 in Table 3 denotes the channel normalization with $\bar{\mu}_c$ and $\bar{\sigma}_c$ computed over the whole utterance. The best overall performance (ALL) is for channel

TABLE 3

DCF Values ($\times 100$) for One-Speaker Detection for Different Handset Conditions with PSS = 0.0

CN type	ALL	SS	SD	DD
1	5.33	2.81	6.78	9.63
2	5.30	2.68	6.35	9.59
3	5.53	3.23	6.89	9.51
0	5.15	2.40	6.32	9.90

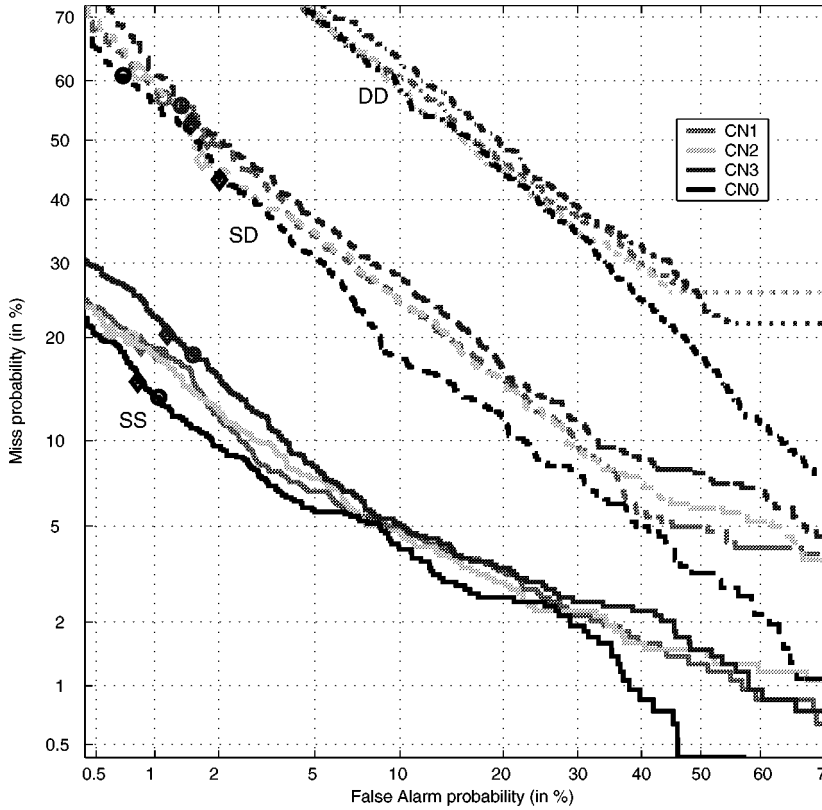

FIG. 3. One-speaker detection performance for different types of channel normalization and PSS = 0.0. DET curves are displayed for the three different channel mismatch factors: SS, SD, and DD. The circles indicate the hard decision points, the minimum DCF is indicated with a diamond.

normalization over the whole utterance (CN0), followed by CN1 and CN2, and the worst performance is for CN3. Thus for one-speaker detection the performance seems to correlate with the length of the window used for computation of the cepstral normalization parameters. However, the impact of different CN types is very small. Figure 3 shows that in other regions of the DET plot (with low miss probability) the superior performance of CN0 is more evident.

Increasing the penalty for short segments (thus creating a bias toward longer segments) has a counterproductive effect on the one-speaker detection performance. Figure 4 shows that no penalty for short segments (PSS = 0.0) is the best setting for all CN types. For higher PSS values the duty cycle (the second factor in Eq. (3)) goes to zero for a larger proportion of the tests (short target segments are no longer detected), concentrating too much of the target and non-target test scores around LLR = 0, and thus reducing the separability between target and non-target score distributions.

Among the systems participating in the 1999 speaker recognition evaluation this system would be ranked in the middle bracket.

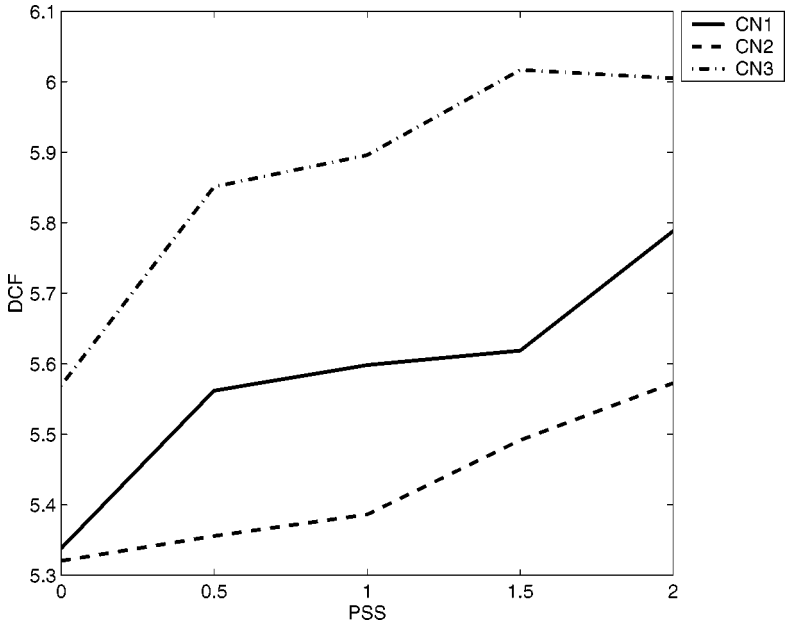

FIG. 4. Impact of the penalty for short segments (PSS) on one-speaker detection performance for different channel normalization types.

3.4. Two-Speaker Detection

The gender mixture of a conversation is likely to have an impact on the performance of two-speaker detection. So results are presented for different (sub)sets of the test data:

ALL: all 15,906 tests;
MF: a subset of 10,538 tests between a male and a female speaker;
FF: a subset of 3674 tests between two female speakers;
MM: a subset of 1694 tests between two male speakers.

Table 4 shows that performance on male–male conversations is best, while female–female conversations produce the highest error figures. However, the crossing lines in Fig. 5 show that the differences in performance between the different gender mixtures depend on the object function that is optimized. The results do not provide unequivocal support for the assumption that mixed

TABLE 4

DCF Values ($\times 100$) for Two Speaker Detection for Different Gender Mixtures with PSS = 0.0

CN type	ALL	MF	FF	MM
1	6.49	6.34	7.20	5.55
2	6.39	6.26	6.66	5.72
3	6.76	6.63	7.10	6.26

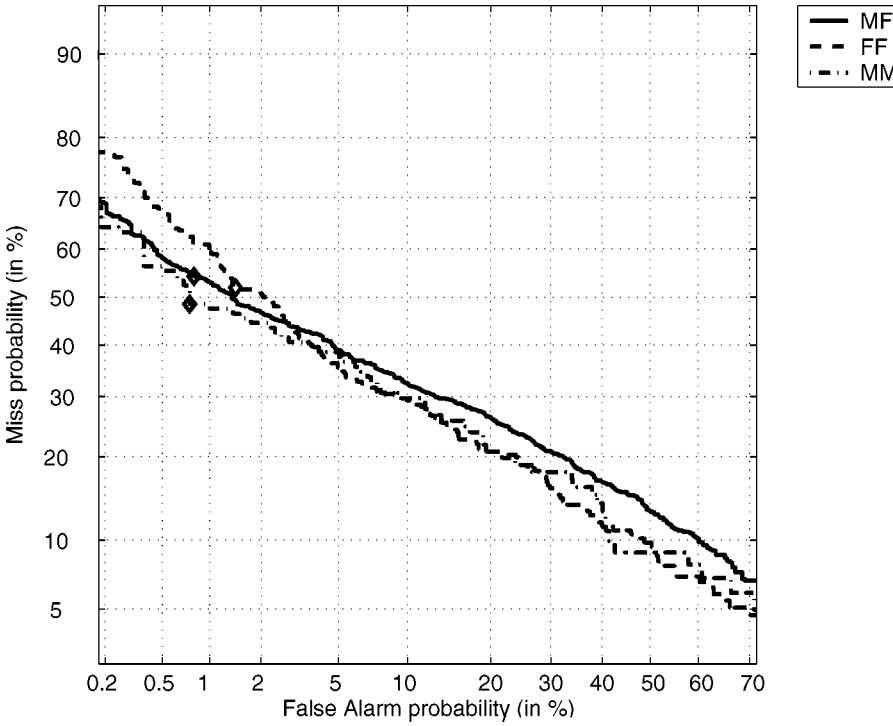

FIG. 5. Two-speaker detection performance for the possible gender mixtures, with channel normalization CN2 and PSS = 0.0. The circles indicate the hard decision points; the minimum DCF is indicated with a diamond.

gender conversations are easier than same gender conversations. At the NIST optimal operation point (with low false alarm rates) it seems that the more females are engaged in the conversation the worse the detection performance is. This may be due to the fact that our system performs worse for females than for males. As a consequence, male-only conversations yield best results.

Channel normalization effects are small for two-speaker detection: overall performance is best for CN1 and CN2 and a bit worse for CN3, which has the shortest time window.

Creation of more global segmentations by increasing the PSS value also has a counterproductive effect on two-speaker detection (Fig. 6). More global segmentations may miss short target segments, at the same time longer segments may be contaminated with the other speaker at the borders of the segments. When all target segments are clustered by integration over T in Eqs. (3) and (4), the integral then misses some short target segments and is contaminated by non-target information, giving a worse LLR estimate than with a detailed segmentation with pure segments. Segments that are too short cause no estimation problems here, because all segments are clustered by the integration over T.

An overall DCF score of 6.39 would place this system among the better ones in two-speaker detection.

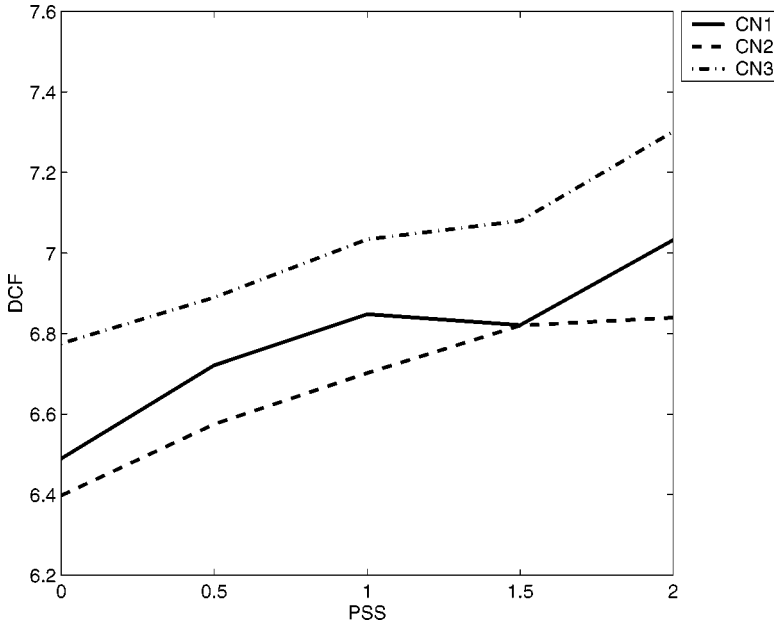

FIG. 6. Impact of the penalty for short segments (PSS) on two-speaker detection performance for different channel normalization types.

3.5. Speaker Tracking

Results for speaker tracking are also presented for different gender mixtures of the conversations. Table 5 shows that it is not necessarily true that mixed gender conversations are easier than same gender conversations. Although one might expect that male–female conversations should have greater interspeaker contrast, there are apparently more important factors than only gender mixture that determine performance. Figure 7 shows that the performance for different gender mixtures also depends on the operation point: around the equal error rate point performance is worst for male–female conversations, and performance is the same for male–male and female–female conversations. We should be careful in drawing conclusions about which task is intrinsically the most difficult.

More consistent are the results for the different channel normalization types (see Fig. 8). CN2 outperforms CN3 (the condition with the shortest sliding window), so having too little data to estimate cepstral mean and standard

TABLE 5

DCF Values ($\times 100$) for Speaker Tracking for Different Gender Mixtures with PSS = 2.0

CN type	ALL	MF	FF	MM
1	8.20	8.15	8.50	7.72
2	7.97	7.93	8.29	7.44
3	8.48	8.43	8.70	8.03

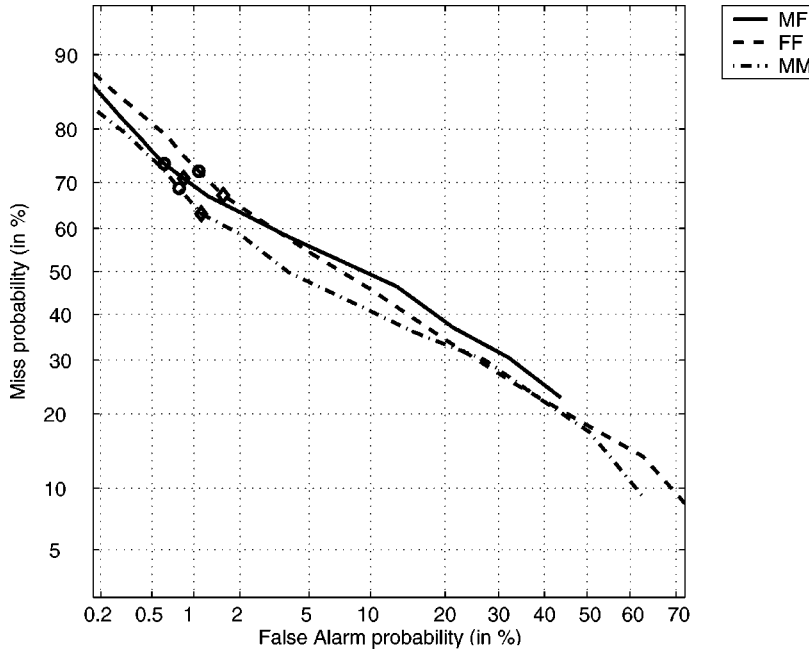

FIG. 7. Speaker tracking performance for the possible gender mixtures, with channel normalization CN2 and PSS = 2.0. The circles indicate the hard decision points; the minimum DCF is indicated with a diamond.

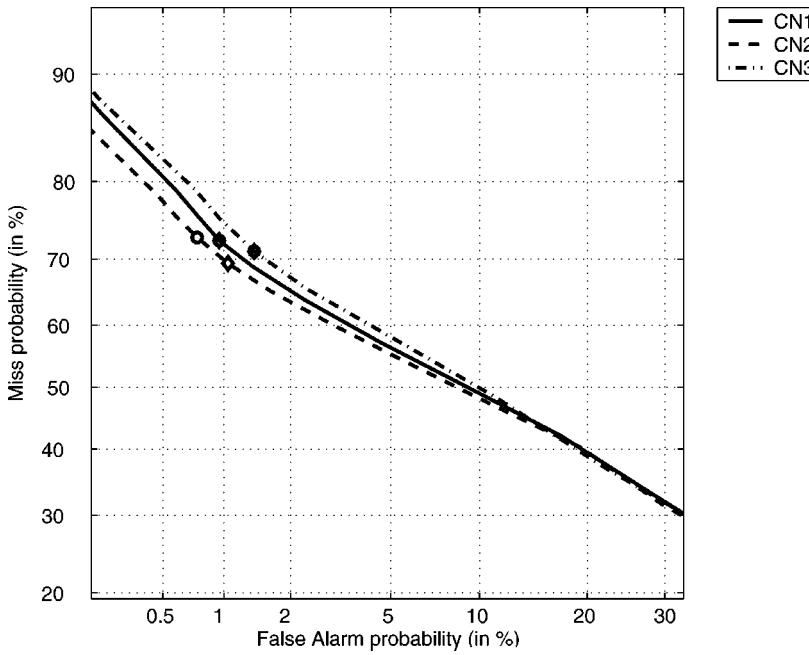

FIG. 8. Speaker tracking performance for different types of channel normalization and PSS = 2.0. The circles indicate the hard decision points; the minimum DCF is indicated with a diamond.

TABLE 6

Mean Segment Length (in seconds) for Different CN Types and PSS Values, Measured over the Segments Assigned to the Target

	PSS				
CN type	0	0.5	1.0	1.5	2.0
1	0.18	0.77	0.79	0.76	0.73
2	0.18	0.84	0.90	0.88	0.84
3	0.16	0.66	0.66	0.62	0.60

Note. The NIST answer key average segment duration is 1.46 s.

deviation harms performance more than the higher probability that the time window contains more than one channel type. The hard decisions in the segmentation needed for CN1 also do more harm than good. This is another case where the maxim "no knowledge is better than false knowledge" applies.

Table 6 proves that increasing the penalty for short segments also increases the mean segment length in the output segmentation from the dynamic programming. The results in Fig. 9 show that speaker tracking performance improves if segmentation becomes more global. The DCF values decrease by about 10% to 7.97 when the penalty for short segments increases from 0.0 to 2.0. This is a major improvement given the fact that the best tracking systems have an overall DCF ranging from 8.5 to 9.5 in the 1999 evaluation [5].

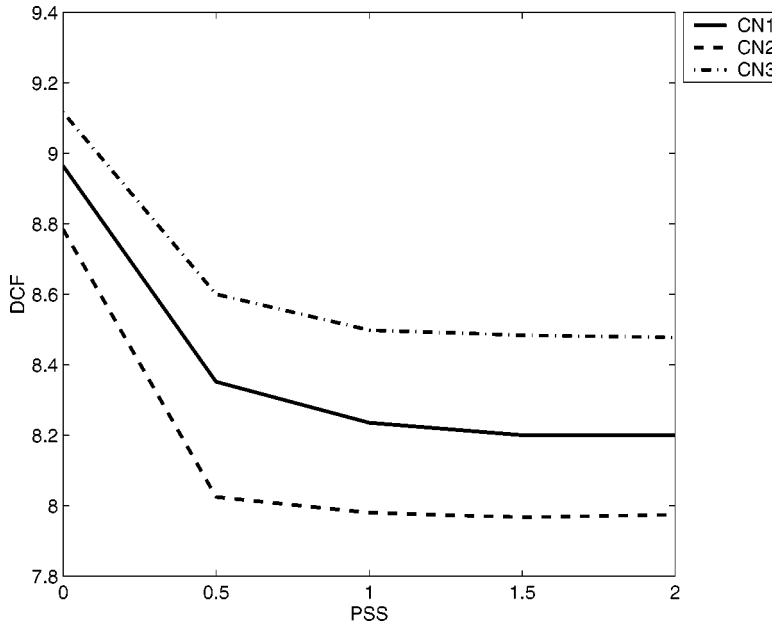

FIG. 9. Impact of the penalty for short segments (PSS) on speaker tracking performance for different channel normalization types.

3.6. Difference between Detection and Tracking

For detection tasks a small penalty for short segments provided best results. However, for speaker tracking best results are obtained with a relatively high penalty for short segments. We conclude that for detection tasks it is most important that segments be homogeneous. Because detection requires a decision per utterance (or conversation), duration of the segments is less important because segment scores are integrated over the collection of target segments to obtain an utterance score. For the tracking task, however, each segment is considered individually and this task requires a decision per frame. In this case homogeneity of segments is less important than segment duration. This duration must be sufficiently long to obtain reliable likelihood estimates. It remains to be clarified whether this difference between detection and tracking tasks is related to the scoring procedures proposed by NIST.

3.7. Tracking Performance Determining Factors

As a side step, it may be of interest to investigate the impact of several conditions on speaker tracking performance. These conditions are (a) speaker balance, where 0 means a "true" dialogue, in which both participants take the floor for approximately the same amount of time, and 1 indicates a monologue where one speaker is talking all the time; (b) percentage of overlapping speech,

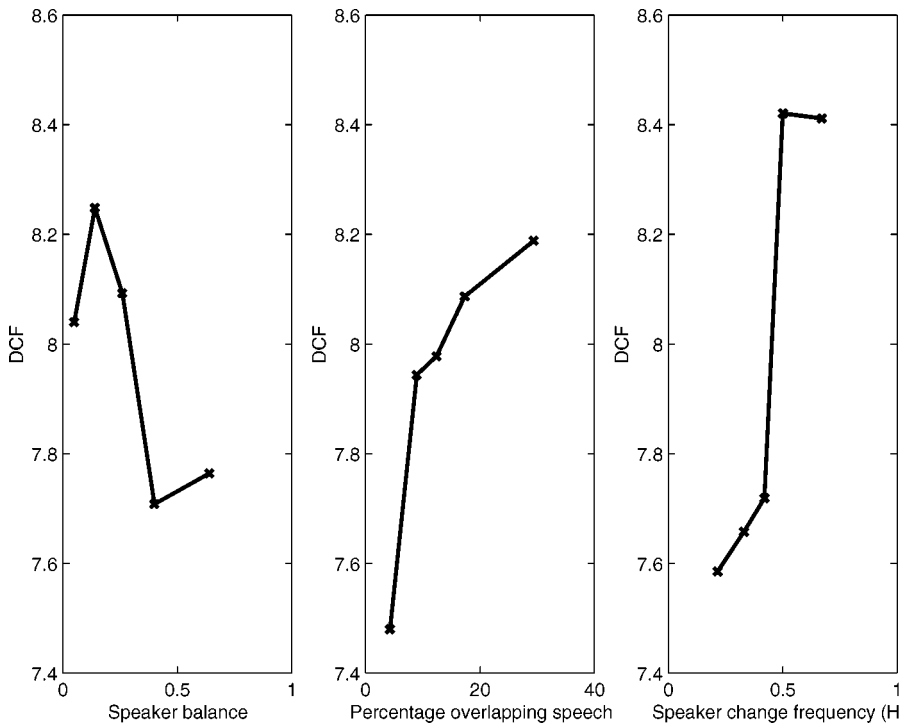

FIG. 10. Factors influencing speaker tracking performance: speaker balance, amount of overlapping speech, and speaker change frequency. Example is for channel normalization CN2 and PSS = 2.0 with an average DCF equal to 7.97.

how often the conversation has two speakers talking at the same time; and (c) speaker change frequency, how many times per second the other speaker takes over the conversation. Note that a speaker change frequency of 0.5 Hz does not mean an average segment length of 2 s since the period between two speaker changes may consist of several segments separated by silence.

Figure 10 shows the impact of these conditions on the detection cost function. It is to be expected that it is easier to detect speakers in monologues than in dialogues. This is also connected to the impact of speaker change: the more frequent speaker change is, the shorter is the period in which one person is speaking, and fewer are the data to estimate normalization parameters and speaker likelihoods; this results in poorer performance for conversations with a high frequency of speaker change. Overlapping speech is clearly a complicating factor, and recognizability of speakers drops critically when the target's speech is masked by speech from another speaker.

4. CONCLUSION

In this paper we compared three different procedures for channel normalization in speaker recognition tasks over the telephone. The differences between the procedures had little impact on the overall performance. However, it seems safe to conclude that a procedure that estimates normalization parameters from sliding windows with a width of 1 s is the best compromise. Longer windows run too high a risk of covering parts of the conversation that originate from different channels. Windows as short as 250 ms apparently suffer from a lack of data, which makes it difficult to obtain meaningful estimates of the normalization parameters. In interpreting these conclusions it must be noted that they may depend to some extent on the specific properties of the tasks. In situations where speaker turns may be expected to be substantially longer than in the switchboard conversations different optimal values may result.

Speaker detection and speaker tracking were performed with a model-based approach: one general silence model, four general speech models (or world models, depending on gender and handset type), and one specific target speaker model available to detect transitions among the target speaker, silence, or another (unknown) speaker. Speaker detection and tracking on conversations require local parameter estimation and local decision making. This paper introduced the concept of delayed decision making and investigated the impact of the level of detail of the segmentation on the likelihood computation and hence on detection and tracking performance.

It was found that one- and two-speaker detection pose requirements different from those of speaker tracking. For detection tasks a small penalty for short segments provided best results under all conditions. However, for speaker tracking best results were obtained with a relatively high penalty for short segments. This is because detection requires a decision on utterance level, whereas tracking requires a decision on frame level.

REFERENCES

1. Chen, S., and Gopalakrishnan, P. S., Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings of the DARPA Workshop,* 1998; available at http://www.itl.nist.gov/iaui/894.01/proc/darpa98/html/bn20/bn20.htm
2. Gish, H., Siu, M.-H., and Rohlicek, R., Segregation of speakers for speech recognition and speaker identification. In *Proceedings of the International Conference On Acoustics, Speech and Signal Processing, Toronto,* 1991, pp. 873–876.
3. Hain, T., Johnson, S. E., Tuerk, A., Woodland, P. C., and Young, S. J., Segment generation and clustering in the htk broadcast news transcription system. In *Proceedings of the DARPA Workshop,* 1998; http://www.nist.gov/speech/proc/darpa98/html/bn30/bn30.htm
4. Koolwaaij, J. W., One speaker detection with the A2RT system. In *Proceedings of the NIST Speaker Recognition Workshop,* 1999.
5. Martin, A., and Przybocki, M., The NIST 1999 speaker recognition evaluation—An overview, *Digital Signal Process.* **10** (2000), 1–18.
6. Melin, H., Koolwaaij, J. W., Lindberg, J., and Bimbot, F., A comparative evaluation of variance flooring techniques on hmm-based speaker verification. In *Proceedings of the International Conference on Spoken Language Processing, Sydney,* 1998, pp. 2379–2382.
7. NIST. 1999 speaker recognition evaluation plan; available at http://www.itl.nist.gov/div894/894.01/spk99/spk99plan.html
8. Przybocki, M. A., and Martin, A. F., The 1999 NIST speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking. In *Proceedings of the European Conference on Speech Technology, Budapest,* 1999, pp. 2215–2218.
9. Reynolds, D. A., The effects of handset variability on speaker recognition performance experiments on the switchboard corpus. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Atlanta,* 1996, pp. 113–116.
10. Viikki, O., and Laurila, K., Cepstral domain segmental feature vector normalization for noise robust speech recognition, *Speech Comm.* **25** (1998), 133–147.