# NEVER-ENDING LEARNING SYSTEM FOR ON-LINE SPEAKER DIARIZATION

*Konstantin Markov*[1,2] *and Satoshi Nakamura*[1,2]

[1]National Institute of Information and Communications Technology, Japan
[2]ATR Spoken Language Communication Research Labs., Japan

## ABSTRACT

In this paper, we describe new high-performance on-line speaker diarization system which works faster than real-time and has very low latency. It consists of several modules including voice activity detection, novel speaker detection, speaker gender and speaker identity classification. All modules share a set of Gaussian mixture models (GMM) representing pause, male and female speakers, and each individual speaker. Initially, there are only three GMMs for pause and two speaker genders, trained in advance from some data. During the speaker diarization process, for each speech segment it is decided whether it comes from a new speaker or from already known speaker. In case of a new speaker, his/her gender is identified, and then, from the corresponding gender GMM, a new GMM is spawned by copying its parameters. This GMM is learned on-line using the speech segment data and from this point it is used to represent the new speaker. All individual speaker models are produced in this way. In the case of an old speaker, s/he is identified and the corresponding GMM is again learned on-line. In order to prevent an unlimited grow of the speaker model number, those models that have not been selected as winners for a long period of time are deleted from the system. This allows the system to be able to perform its task indefinitely in addition to being capable of self-organization, i.e. unsupervised adaptive learning, and preservation of the learned knowledge, i.e. speakers. Such functionalities are attributed to the so called *Never-Ending Learning systems*. For evaluation, we used part of the TC-STAR database consisting of European Parliament Plenary speeches. The results show that this system achieves a speaker diarization error rate of 4.6% with latency of at most 3 seconds.

***Index Terms***— Speaker diarization, Speaker segmentation, On-line GMM learning, Never-ending learning.

## 1. INTRODUCTION

The task of efficient and effective automatic indexing and searching of the growing volumes of recorded spoken documents, such as broadcasts, voice mails, meetings and others, requires human language technologies that can not only transcribe spe-ech, but can also extract different kinds of non-linguistic information. This information, often called meta-data, includes speaker turns, channel changes, and others. Identifying and labeling the sound sources within a spoken document is the task of audio diarization. A main part of the audio diarization process is the speaker diarization or speaker segmentation and clustering. In other words, it is the task to find out "who spoke when".

Speaker diarization is currently the focus of the most efforts in the audio diarization research, which has been also driven by the recent NIST Rich Transcription [1] and Speaker Recognition [2] evaluations. Broadcast news audio, meetings recordings or telephone conversations are one of the main domains for speaker diarization research and development. In some cases, prior information about the task can be available. This may be an example speech from speakers of a meeting or from the main anchors of a broadcast. However, from a system portability point of view, it is better to use less or no prior knowledge at all.

Most of the current speaker diarization systems perform several key sub-tasks which are: Speech detection, Speaker change detection, Gender classification and Speaker clustering [3]. To improve the performance, in some cases, cluster recombination and re-segmentation are also used [4]. The speech detection is aimed to find those regions of the audio which consist of speech only. The most popular technique to perform this task is the maximum-likelihood classification with Gaussian mixture models (GMM). They are usually trained in advance from some labeled data and, in the simplest case, there are only two models for speech and non-speech data [5]. Some systems use several models depending on the speaker gender and the channel type [6, 7]. Another approach that has been found useful is to perform a single or multi-pass Viterbi segmentation of the audio stream [8, 9]. For the broadcast news data, the typical speech detection error rates are 2% - 3%. After the speech segments are identified, speaker change detection is used to find out any possible speaker change within every segment. If such is detected, the segment is further split into smaller segments each of which belongs to a single speaker. There are two main techniques for change detection. The first one finds potential change point in a window by determining whether it is better modeled by two rather than one distribution using the Bayesian information criterion (BIC) [8]. The second one is based on measuring the distance, Gaussian divergence [10] or

generalized likelihood ratio [11], between two fixed length windows represented most often by a single Gaussian. A distance peak that is above certain threshold is then considered as a change point. The problem is that single Gaussian function is a rough model of the data distribution from one segment with typical length of 2-5 seconds. This inevitably introduces detection errors and increases the error rate sensitivity to the decision threshold. The gender classification is used to split the segments into two groups (male and female) which reduces the load of the next clustering task as well as to give more information about the speakers. Typically, two GMMs, one for each gender, are trained in advance and maximum-likelihood is used as decision criterion. Reported gender classification error rates are usually 1%-2% [3]. The last sub-task, the speaker clustering, is to assign each segment with its correct speaker label. This is done by clustering segments into sets corresponding to speakers. The most widely used approach is hierarchical, agglomerative clustering with BIC stopping criterion [9, 12]. Each cluster is usually represented by a single Gaussian and the generalized likelihood ratio (GLR) [13] has been commonly used as between clusters distance measure. Variations of this method have also been proposed [7, 14], but they are still based on the same bottom-up clustering technique. Although, quite successful, agglomerative clustering approach has several drawbacks that limit the potential use of the speaker diarization systems in the real-world, real-time applications. First, it requires all the speech segments to be available before the clustering starts and, therefore, makes on-line processing impossible. Second, the computational load increases almost exponentially with the number of segments [15]. Finally, the performance is greatly affected by the stopping criterion which is considered as a critical part of the algorithm [3]. A sequential algorithm based on the leader-follower clustering [16] and suitable for on-line operation has been proposed recently [15]. However, as in the agglomerative clustering method, the speech segments are modeled by a single Gaussian distribution and the GLR is used as a distance metric. This reduces the clustering accuracy for short segments and delays the decision until the whole segment is received. In consequence, the system latency becomes dependent on the segment's length which can be up to 30 sec. or even longer. Another sequential technique where speakers are represented by subspaces has also been studied [17]. However, it requires at least 5 sec. long speech segments and has high miss and false alarm rates.

In this paper, we propose a new speaker diarization system, which in contrast to those mentioned above operates on-line, in less than real time, and has low latency of up to few seconds. It performs all the sub-tasks of a standard diarization system, but the within segment speaker change detection. Based on the observation that most speaker change points occur during the non-speech regions [7], we assume that each speech segment belongs to a single speaker. In case this assumption cannot be justified, our system can be easily upgraded with a speaker change detection module. What makes it significantly different from the other systems is the way the segment clustering is performed as well as the overall operating algorithm, which is based on the Never-Ending Learning (NEL) principle [18]. In our system, when assigning speaker label to a given segment, first, it is decided whether it belongs to one of the known speakers or to a new speaker. Then, in the former case, speaker identification is performed and the winning speaker label is assigned to the segment. In the latter case, new speaker is registered to the system and his/her model is created. This is similar to the classical open-set speaker identification task. Each speaker is represented by a GMM which is learned on-line every time it has been a winner. New speaker's GMM is created by spawning the corresponding gender GMM. In addition, each speaker GMM has a time counter which is set to zero whenever it wins the identification. Otherwise, the counter is incremented by the current segment length. Models whose counter reaches some threshold T, are deleted from the system. This way, the system can operate indefinitely, adapting itself to the environment changes, i.e. changes in the number of speakers and their characteristics, and acquiring new knowledge, i.e. new speakers, in an unsupervised manner without catastrophic forgetting (i.e. newly learned knowledge does not wipe out previously learned one). Such systems are called *Never-Ending Learning systems*.

Next section describes our system in details. Evaluation setup and the experimental results are presented in Section 3. Conclusions and future work are summarized in the last section.

## 2. SYSTEM DESCRIPTION

### 2.1. Overview

The block diagram of the speaker diarization system is shown in Fig.1. It consists of several modules and a set of GMMs. There are one pause GMM, two gender dependent GMMs and variable number of speaker GMMs. Block arrows show how modules share these models. Thin arrows show the control flow. Unsegmented audio data is fed to the voice activity detection module. It outputs speech segments start and end points. As soon as the start point is decided, frame by frame, the likelihoods from all the GMMs (except the pause GMM) are accumulated for some time in the novelty detection module. This time, called decision time (DT), is essentially the system latency time. Then based on the accumulated likelihoods, it is decided whether the segment belongs to an old speaker or not. If it is a new speaker, its gender is determined in the gender identification module using the accumulated likelihoods from the two gender GMMs. Then, from the corresponding gender GMM, a new GMM is spawned by copying its parameters in the new model generation module. This GMM is given new speaker name and is inserted in the

system speaker GMM set. The new model is learned on-line using the speech data from the start point until some time, called learning time (LT). In case LT is bigger than the current segment's length, it is set to this length, but only for the current segment. The same holds for the DT. When the novelty detection module decides that segment belongs to an old speaker, this speaker is identified in the speaker identification module by the maximum likelihood criterion. Each speech segment is labeled by the name of the winning speaker, either new or old.

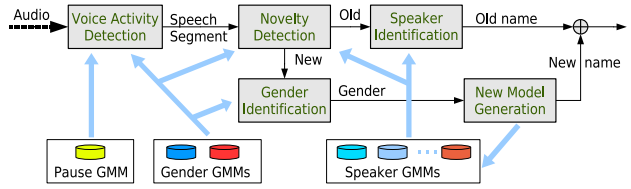


**Fig. 1**. Block diagram of the system. Block arrows show how modules share GMMs. Thin arrows show the control flow.

The system works on-line and and its operation is schematically shown in Fig. 2. The speech segments and their reference speaker labels are at the top of the figure. The bottom part shows the speaker models and how they change in time. For each speech segment, there is a winning model indicated by a thick border line. At the beginning, there are only three GMMs: one for pause (not shown for clarity) and two for each speaker gender. They are trained in advance from some labeled data. For the first segment, the speaker gender is identified (male in the figure) and a new GMM is created from the male GMM. It is learned on-line with the segment's data, and from this point it becomes the GMM for Speaker 1 (SP1 in the figure). The next segment is from the same speaker, so the SP1 GMM will be the winner. It is again learned on-line with the second segment's data. The third segment comes from a female speaker and the same procedure is repeated resulting in a set of two speaker GMMs. This way, the system generates a set of speaker models on the fly. If some GMM (SP1 in the figure) has not been a winner for a long time, it is deleted from the system (indicated by an "X" on the figure). Such operating mode allows the system to work indefinitely.

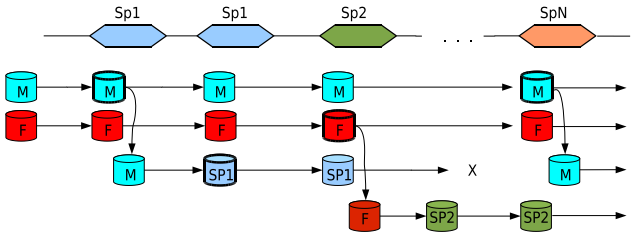


**Fig. 2**. System operation. For each speech segment, the winning GMM is denoted by bold border lines. The pause GMM is not shown for clarity.

### 2.2. Voice activity detection

For the voice activity detection (VAD), we use the standard model based approach. Non-speech events (pauses in this case, but other event can also be modeled) are represented by a single GMM and the speech is modeled by the two gender dependent GMMs. For each frame, the non-speech and speech (the better one from the two GMMS) likelihoods are passed through two separate median filters and the frame's label (speech / non-speech) is assigned by comparing the filters output. Then, a simple logic decides segments start and end points taking into account such requirements as minimum segment length (MSL), maximum pause in segment (MPS) and maximum speech in pause (MSP).

### 2.3. Gender identification

The gender identification module uses the same gender dependent GMMs as the VAD module. Frame likelihoods calculated already during the voice activity detection are accumulated from the segment's start point for a time set by the DT parameter. Then, the speaker gender is determined by a simple maximum-likelihood classification.

### 2.4. Novelty detection

The purpose of this step is to decide whether the current segment comes from one of the registered system's speakers or from a new speaker. This is a typical hypothesis testing problem, where the standard solution is the likelihood ratio test. It is formulated as follows:

$$X \in \begin{cases} \omega_0, & \text{if } \mathrm{L(X)} > \theta \\ \omega_1, & \text{if } \mathrm{L(X)} < \theta \end{cases} \quad (1)$$

where $X = \{x_i\}, i = 1, \ldots, DL$ is a decision length speech segment, $\omega_0$ is a class corresponding to the hypothesis $H_0$, i.e. old speaker. Respectively, $\omega_1$ corresponds to $H_1$, i.e. new speaker. The likelihood ratio is:

$$L(X) = \frac{p(X|\omega_0)}{p(X|\omega_1)} \quad (2)$$

There are various ways to define $p(X|\omega_i)$. Considering the available set of GMMs, a straightforward approach is to define them as:

$$\begin{aligned} p(X|\omega_0) &= P_{sp} = \max_{\lambda_j \in \Lambda} p(X|\lambda_j) \quad (3) \\ p(X|\omega_1) &= P_{gen} = \max(p(X|\lambda_{male}), p(X|\lambda_{female})) \end{aligned}$$

where $\Lambda = \{\lambda_j\}$ is the current set of speaker GMMs. Another approach, often used in speaker verification is to define $p(X|\omega_1)$ as:

$$p(X|\omega_1) = P_{ave} = \frac{1}{n-1}(\sum_j p(X|\lambda_j) - P_{sp}) \quad (4)$$

i.e. the average of all model likelihoods except for the winning model. Here $n = |\Lambda|$ is the size of the speaker set. Experimentally we verified that combining the two approaches works better than either of them. In this case the likelihood ratio is:

$$L(X) = \frac{P_{sp}^2}{P_{gen} P_{ave}} \tag{5}$$

The threshold $\theta$ is usually estimated using a development data set.

Although separated in a different module, the speaker identification is implicitly performed during the novelty detection task since the best speaker likelihood is required for the likelihood ratio calculation. The same holds for the gender identification. If the winning hypothesis is $H_0$, then the best speaker is identified from $P_{sp}$. Otherwise, the winning gender is found from $P_{gen}$.

### 2.5. On-line GMM learning

This step is the one that allows the whole system to operate on-line and makes it different from all other systems. The main algorithm for off-line GMM parameter estimation is the Expectation-Maximization (EM) algorithm. Not long ago, incremental versions of it were proposed [19, 20], which facilitated the development of on-line variants [21, 22]. In the on-line EM, statistics and parameters are updated after each observation $x$ using the following equations:

$$\begin{aligned} \ll f(x,y) \gg_i (t) = &\ll f(x,y) \gg_i (t-1) + \\ &\eta(t)[f(x(t),y(t))P_i(t) - \ll f(x,y) \gg_i (t-1)] \end{aligned} \tag{6}$$

where $\ll f(x,y) \gg_i (t)$ is the statistic function of the complete data $(x,y)$. The posterior probability of the Gaussian component $i$ given the previous parameter set $\Theta_{t-1}$ is defined as $P_i(t) \doteq P(i|x(t), y(t), \Theta_{t-1})$. The learning rate $\eta(t)$ satisfies the constraints:

$$1 \geq \eta(t) \geq 1/t \tag{7}$$

The new parameters $\Theta_t$ are obtained from:

$$\begin{aligned} c_i(t) &= \ll 1 \gg_i (t) \\ \mu_i(t) &= \ll x \gg_i (t) / \ll 1 \gg_i (t) \\ \sigma_i^2(t) &= \ll x^2 \gg_i (t) / \ll 1 \gg_i (t) - \mu_i^2(t) \end{aligned} \tag{8}$$

The on-line EM converges faster than the standard EM, but even few iterations could increase too much the computational load for a real-time system. On the other hand, given an infinite number of data drawn from the same distribution, the on-line EM can be considered as a stochastic approximation [23]. In practice, this means that as long as there is enough data, model parameters can be approximated in one pass. In this case, the learning rate $\eta(t)$ should satisfy the conditions:

$$\eta(t) \stackrel{t \to \infty}{\longrightarrow} 0, \ \sum_{t=1}^{\infty} \eta(t) = \infty, \ \sum_{t=1}^{\infty} \eta^2(t) < \infty \tag{9}$$

Commonly used function that satisfies these conditions as well as Eq.(7) is:

$$\eta(t) = \frac{1}{at + b} \quad 1 > a > 0 \tag{10}$$

where $a$ and $b$ are parameters which control the learning process. The past samples forgetting speed depends on $a$, while $b$ sets the learning speed of the new samples.

This algorithm allows fast and inexpensive on-line learning of the system GMMs. As in the batch EM case, the initial parameter values play important role in the learning speed and the precision of the final estimates. Therefore, it is desirable for the initial values to be as close as possible to the true ones. In our system, the gender dependent GMM parameters are the best available initial values for every speaker model and that is why they are used for the new GMM generation.

## 3. EXPERIMENTS

### 3.1. Database and pre-processing

For the system evaluation, we used the data released for the TC-STAR 2007 evaluation campaign [24]. The data consists of recordings of the European Parliament plenary speeches. From the training part of the database, we selected about 20 min of silence data for building the pause model. For the gender dependent models, about 2 min. of speech from each of 20 male and 15 female speakers was used. The official development set was used as development data, and the evaluation set from the TC-STAR 2006 campaign was used for the final system evaluation.

All audio data were transformed into 26 dimensional feature vectors consisting of 12 MFCC coefficients, power and their first derivatives. The frame length and rate were 20 and 10 ms. respectively.

### 3.2. Preliminary experiments

Before running the on-line experiments, we investigated the performance of the on-line learning algorithm in separate off-line tests. First, using the data selected for the gender models, we trained off-line one GMM for every speaker. This allows us to compare the on-line and off-line learning algorithms' speaker identification performance. For tests, we used about 30 sec. of each speaker's data, but different from that used for training. Then we run two types of experiments. One is a speaker identification with the off-line trained GMMs. In the other, each speaker GMM was replaced by its on-line learned version, one at a time, and the results were averaged over all speakers. Table 1 shows the identification rates for the test data of different lengths when the on-line learning is done using 2 or 4 sec. of data. The size of the GMMs in these experiments was 64, and the on-line learning parameters were set to $a = 0.999$ and $b = 1000$, which were found to give the

**Table 1**. Speaker identification performance for on-line and off-line trained GMM (%).

| Test length | On-line learning time | | Off-line learning |
|---|---|---|---|
| | 2 sec. | 4 sec. | |
| 1 sec. | 95.6 | 96.6 | 98.5 |
| 2 sec. | 98.8 | 99.6 | 99.7 |
| 3 sec. | 99.5 | 100 | 100 |

best performance. The results show that the on-line learning can produce models pretty close to those trained off-line even with small amount of learning data.

We also checked the gender identification performance using the same test data. It was 97.3%, 98.4% and 99.6% for the test data of 1, 2 and 3 seconds respectively. Actually, the results are little bit biased because the test speakers are those used for the gender GMMs training. Nevertheless, we don't expect significant drop in the performance for the real system.

### 3.3. On-line experiments

In these experiments, we first evaluated the performance of the voice activity detector. The evaluation metric was the speaker diarization error rate (DER) given that all speech segments have correct speaker label. The DER is a time wighted sum of miss errors, false alarms and speaker errors. Since there will be no speaker errors in this setup, the DER will show the VAD performance and it is shown in Table 2 for both the development "dev" and evaluation "eval" data. The minimum segment length (MSL) was set to 1 or 2 seconds. Bigger values did not improve the results. Typically, a forgiveness collar of 0.25 sec around the reference segment boundaries is set when the DER is calculated. Results with no collar are also presented in the table.

**Table 2**. VAD performance in terms of DER (%).

| Min. segment length | Collar = 0.0 | | Collar = 0.25 | |
|---|---|---|---|---|
| | dev | eval | dev | eval |
| 1 sec. | 4.3 | 4.5 | 1.9 | 2.5 |
| 2 sec. | 4.5 | 4.6 | 2.3 | 2.5 |

In the next experiments, we tested the speaker segmentation performance, where the main parameter to be determined was the novelty detection threshold. For that, we used the development data only, and the true segment boundaries. This way, the DER will show only the speaker errors. The results when the maximum decision length (DL) was varied from 1 to 5 seconds, are shown in Fig.3. Here, the on-line learning time (LT) was set to 10 seconds. Bigger LT, or even using the whole segments for learning, did not improve the performance, but only increased the computational load. As the

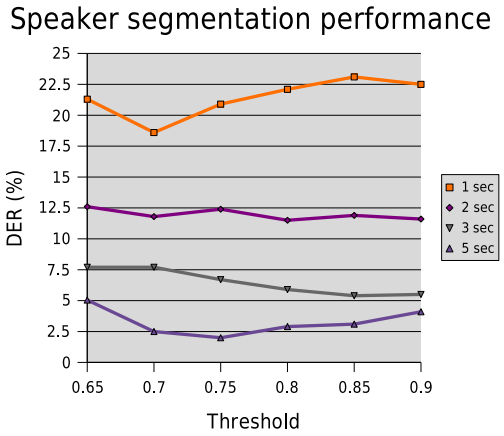


**Fig. 3**. Speaker segmentation performance in terms of DER for different novelty detection threshold values.

figure shows, the novelty detection is quite insensitive with respect to the threshold.

For the whole speaker diarization system evaluation, we set the novelty detection threshold to 0.8 and the DER results for both the development and evaluation data are summarized in Table 3. As can be seen, the performance improves rapidly

**Table 3**. The full system performance in terms of DER (%).

| System latency | Collar = 0.0 | | Collar = 0.25 | |
|---|---|---|---|---|
| | dev | eval | dev | eval |
| 1 sec. | 14.1 | 21.2 | 11.5 | 19.4 |
| 2 sec. | 9.4 | 18.8 | 6.7 | 16.8 |
| 3 sec. | 7.2 | 13.8 | 4.6 | 11.9 |
| 4 sec. | 6.6 | 13.1 | 4.0 | 11.3 |
| 5 sec. | 6.6 | 12.1 | 3.9 | 10.2 |

when the maximum DL, i.e. the system latency, is increased to $3 \sim 4$ sec. and then stays almost the same. The error rates for the evaluation data are about two times higher than the development data, which suggests that the DER is sensitive to the irrecoverable errors inherent in the on-line, one-pass systems. Nevertheless, the overall performance is less than 10%, which is in the range of the best off-line multi-pass speaker diarization systems. As for the processing speed, the system showed real time factor of less than 0.1xRT.

## 4. CONCLUSION AND FEATURE WORK

We described a new speaker diarization system that works on-line, faster than real-time, and has high performance. The system consists of several modules, each of which is based on conventional methods, but the system design and the usage of the on-line EM for GMM learning allowed to achieve some

unique capabilities, such as infinite operation, self-organization and knowledge preservation.

The system was recently built and this was its first evaluation, so there is a lot of room for further improvement in each of the modules. Especially challenging is to develop an algorithm which would prevent the propagation of the previously made errors.

## 5. REFERENCES


[1] NIST, "Benchmark Tests: Rich Transcription (RT)," Online: http://www.nist.gov/speech/tests/rt/.

[2] NIST, "Benchmark Tests: Speaker Recognition," Online: http://www.nist.gov/speech/tests/spk/.

[3] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Trans. ASLP*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[4] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving Speaker Diarization," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Nov. 2004.

[5] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Toward Robust Speaker Segmentation: The ICSI-SRI Fall 2004 Diarization System," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Nov. 2004.

[6] P. Nguyen, L. Rigazio, Y. Moh, and J.-C. Junqua, "Rich transcription 2002 site report: Panasonic speech technology laboratory (PSTL)," in *Proc. Rich Transcription Workshop (RT-02)*, 2002.

[7] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proc. Eurospeech*, Sept. 1999, pp. 1031–1034.

[8] D. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Nov. 2004.

[9] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu, "The LIMSI 2006 TC-STAR transcription system," in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 123–128.

[10] M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news," in *Proc. DARPA Speech Recognition Workshop*, Feb. 1997, pp. 97–99.

[11] A. Rosenberg, A. Gorin, Z. Liu, and S. Parthasarathy, "Unsupervised speaker segmentation of telephone conversations," in *Proc. ICSLP*, Sept. 2002, pp. 565–568.

[12] F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, D. Pineda, D. Seppi, and G. Stemmer, "The ITC-irst transcription systems for the TC-STAR-06 evaluation campaign," in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 117–122.

[13] S. Stuker, C. Fugen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y.-C. Tam, and M. Wofel, "The ISL TC-STAR Spring 2006 ASR evaluation systems," in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 139–144.

[14] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs," in *Proc. ICSLP*, Oct. 2004, pp. 2329–2332.

[15] D. Liu and F. Kubala, "Online Speaker Clustering," in *Proc. ICASSP*, May 2004, pp. 333–336.

[16] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, Inc., Second edition, 2001.

[17] M. Nishida and Y. Ariki, "Real time speaker indexing based on subspace method - Application to TV news articles and debate," in *Proc. ICSLP*, Dec. 1998, vol. 4, pp. 1347–1350.

[18] K. Markov and S. Nakamura, "Never-Ending Learning with Dynamic Hidden Markov Network," in *Proc. INTERSPEECH*, Aug. 2007, to be published.

[19] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse and other variants," in *Learning in Graphical Models*, M. Jordan, Ed., pp. 355–368. The MIT Press, 1999.

[20] S. Nowlan, *Soft competitive adaptation: Neural Network learning algorithms based on fitting statistical mixtures*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1991.

[21] M. Sato and S. Ishii, "On-line EM algorithm for the Normalized Gaussian Network," *Neural Computation*, vol. 12, pp. 407–432, 2000.

[22] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 275 – 300, May 2004.

[23] H. Kushner and G. Yin, *Stochastic approximation algorithms and applications*, Springer-Verlag, New York, 1997.

[24] TC-STAR, "Technology and Corpora for Speech to Speech Translation," Online: http://www.tc-star.org/.