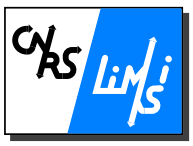
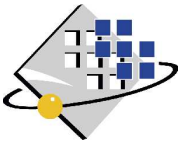
Structuration automatique en locuteurs par approche acoustique

Xuan Zhu





N. d'ordre

Université Paris XI
UFR scientifique d'Orsay

Thèse présentée pour obtenir

Le grade de docteur en sciences de l'université Paris XI Orsay

par Xuan Zhu

Sujet : Acoustic-based Speaker Diarization

Soutenue le 15 Octobre 2007 devant la commision d'examen :

M. Xavier Rodet	Président
M. Jean-François Bonastre	Rapporteur
M. Christian Wellekens	Examinateur
M. Sylvain Meignier	Examinateur
M. Jean-Luc Gauvain	Directeur de thèse
M. Claude Barras	Encadrant de thèse

Abstract

This thesis presents a work focusing on the topic of speaker diarization for different types of audio recordings, especially including broadcast news (BN) and meetings. The speaker diarization is a relatively recent speech processing technique, but it has attracted strong research efforts due to its great benefit to other speech technologies, such as rich transcription, audio indexing and speaker recognition.

The speaker diarization task aims to answer the question of "who spoke when" for a given audio stream. A diarization system is required to produce a sequence of speaker turns with their corresponding speaker identities. This thesis work is carried out following the assumption that no a priori knowledge of the speakers voice or the number of speakers is available.

The proposed BN speaker diarization system is developed from a data partitioning system that was designed as a pre-processing stage to automatic transcription system. The main modification is to combine two speaker clustering stages, where a Bayesian Information Criterion (BIC) clustering using single full-covariance Gaussian models is performed to provide a under-clustering and the resulting clusters are recombined via a second clustering stage relying on Gaussian Mixture Model (GMM) based speaker identification techniques. The implemented BN speaker diarization system has been examined in both the international NIST Rich Transcription 2004 Fall (RT-04F) evaluation and a French Technolangue ESTER evaluation on broadcast news data and provided the state-of-the-art diarization performance in both evaluations.

As there has been strong interest in the meeting domain recently, some adaption on the speaker diarization system have been implemented from broadcast news to meetings, within the framework of the European project CHIL (Computers in the Human Interaction Loop). The adapted diarization system integrates a new speech activity detector based on log-likelihood ratio. Various feature normalization techniques and different sets of acoustic features are also explored by the adapted system. This meeting diarization system is found to provide comparable performance on different sub-types of meeting data (i.e conference and lecture room meetings). In the last NIST RT meeting recognition evaluation, the adapted diarization system had an overlap diarization error of 26% approxiately on the conference and lecture test data.

The detection of overlapping speech is preliminarily studied on telephone speech. Although the proposed detection algorithm has not yet given enough detection performance to be integrated into the speaker diarization system, it provides a starting point for future researches.

Remerciements

Je tiens tout d'abord à remercier les membres de mon jury pour leur participation à la soutenance de ma thèse: Monsieur Xavier Rodet pour avoir présidé le jury, Monsieur Jean-François Bonastre, rapporteur de cette thèse, pour avoir consacré du temps à la lecture de ce document et pour ses remarques constructives, ainsi que Monsieur Christian Wellekens et Monsieur Sylvain Meignier pour leur participation active au jury et l'intêret qu'ils ont porté à ces travaux.

Je remercie également Monsieur Chuck Wooters qui a accepté d'être rapporteur de cette thèse malgré une période chargée à cause de son changement de poste.

Je tiens à remercier Monsieur Jean-Luc Gauvain, directeur de cette thèse, qui, malgré sa lourde tâche de directeur du group Traitement du Langage Parlé au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, a suivi ce travail au long de ces trois années.

Je voudrais exprimer toute ma gratitude à Monsieur Claude Barras qui a largement contribué à l'accomplissment de ce travail. En tant que co-directeur de cette thèse, Claude a efficacement dirigé mes recherches et m'a donné de nombreux conseils pour aboutir cette thèse. Je le remercie également pour sa disponibilité, sa bonne humeur et ses mille gentillesses.

Je remercie aussi tous les membres de TLP pour leur contribution aux conditions de travail chaleureuses. J'adresse de vifs remerciements à Madame Lori Lamel pour ses relectures de nombreux articles et ses aides infaillibles.

J'apprécie tous les petits doctorants du groupe TLP, particulièrement les trois filles: Bianca, Laurence et Cecile, qui m'ont supportée et qui m'ont énormément aidée pendant mes séjours en France pour cette thèse. Je souhaite une longue carrière dans les recherches de la parole et une vie personnelle très heureuse à chacun des ces amis.

Enfin, je remercie mes parents et mon mari pour leur amour et leur soutien tout au long de cette thèse.

Contents

List of Figures

List of Tables

Part I

Résumé de la thèse en Français

Introduction

Cette thèse présente un travail portant sur la structuration en locuteurs pour différents types d'enregistrements audio: journaux d'actualités télévisés ou radiophoniques et réunions. La structuration en locuteurs est une technique relativement récente du traitement de la parole, mais elle a suscité un fort intérêt de recherche en raison de ses applications dans d'autres technologies de la parole, telles que la transcription enrichie, l'indexation audio et l'identification du locuteur.

La structuration en locuteurs, également appelée segmentation et regroupement du locuteurs, est un processus qui consiste à diviser un flux audio en segments homogène en fonction de l'identité du locuteur. C'est un aspect de la structuration de l'audio, avec la catégorisation de la musique, du bruit de fond ou du canal de transmission. La structuration en locuteurs vise à répondre à la question de "qui a parlé quand" pour un document sonore donné. Un système de structuration en locuteurs doit ainsi produire une séquence de segments de parole dans laquelle chaque segment est borné par des instants de changement de locuteur ou des transitions entre parole et non-parole et marqué avec l'identité du locuteur impliqué.

Deux applications principales de la structuration en locuteurs peuvent être trouvées dans la littérature. D'une part, la structuration en locuteurs est une étape très utile de prétraitement pour les systèmes de reconnaissance automatiques de la parole. En séparant les segments de parole des autres, le système de reconnaissance a seulement besoin de traiter les segments audio contenant de la parole, réduisant de ce fait le temps de calcul. En groupant des segments de même nature acoustique, des modèles spécifiques à cette condition peuvent être employés pour améliorer la qualité de la reconnaissance. En groupant des segments du même locuteur, la quantité de données disponible pour l'adaptation non supervisée des modèles au locuteur est augmentée, ce qui peut améliorer significativement la performance en transcription. D'autre part, la structuration en locuteurs est une technique très utile pour les systèmes de transcription enrichie et d'indexation par locuteur. Cette information peut être utilisée pour améliorer la lisibilité d'une transcription automatique en structurant un document sonore par locuteur et dans certains cas en fournissant l'identité du locuteur. Elle peut également être intéressante pour l'indexation des documents multimédia.

Il y a trois domaines principaux concernés par les recherches sur la structuration en locuteurs :

- **Les conversations téléphoniques** impliquant deux locuteurs, chaque locuteur étant enregistré sur un canal individuel. La tâche de structuration en locuteurs a été évaluée pour la première fois sur ce type de données dans les évaluations d'identification du locuteur organisées par National Institute of Standards and Technology (NIST) de 2000 à 2002 [NISTSRE].

- **Les journaux télévisés ou radiophoniques** incluant des conditions acoustiques plus complexes comme la parole de studio et téléphonique, de la musique seule, de la musique avec parole superposée etc. ainsi que de multiples locuteurs présents dans les documents. Ce genre de données est devenu le domaine principal de recherches pour la structuration en

locuteurs de 2002 à 2004 dans les campagnes d'évaluation NIST en transcription enrichie [NIST, 2003; NIST, 2004].

- **Les réunions** divisées en sous-catégories comme les conférences et les séminaires, où la parole est complètement spontanée avec de fréquentes zones de parole superposé et où les signaux présentent des déformations importantes dans un environnement bruyant en raison de l'utilisation de microphones éloignés. Ce domaine a suscité plus d'attention en recherche dans les évaluations sur la structuration et la transcription automatique de réunions organisées par le NIST de 2004 à 2007 [NIST, 2005; NIST, 2006; NIST, 2007], avec la coopération des projets européens CHIL (Computers in the Human Interaction Loop) and AMI (Augmented Multiparty Interaction).

Les trois domaines décrits ci-dessus présentent des difficultés différentes pour la structuration en locuteur, en raison des différences dans le nombre de locuteur, la qualité du signal, les types de sources acoustiques, les durées des tours de parole et le style de discours. Généralement, des systèmes de structuration en locuteurs sont développés pour un domaine spécifique. Plus récemment, un certain travail de recherches a été effectué pour développer des systèmes de structuration ayant une meilleure portabilité entre différents domaines [Anguera *et al.*, 2006c; Wooters and Huijbregts, 2007].

Bien qu'il soit possible d'avoir également des connaissances a priori dans certaines applications spécifiques (par exemple, la connaissance préalable de la voix des présentateurs habituels sur des stations de télévision ou de radio particulières), cette thèse adoptera la définition suivante: aucune échantillon de voix des locuteurs ou même du nombre de locuteurs est disponible. C'est la définition utilisée pour la tâche de structuration en locuteurs dans les évaluations du NIST depuis 2000 [NIST, 2000]. Sous cette hypothèse et en considérant que la structuration en locuteurs est une tâche relative à un enregistrement audio isolé, seules des identités de locuteurs relatives, internes à l'enregistrement, sont produites par le système de structuration.

Architecture générale d'un système de structuration en locuteurs

Les techniques actuellement utilisées pour la structuration en locuteurs dépendent en grande partie des méthodes statistiques et probabilistes employées pour modéliser les différences des caractéristiques acoustiques. La plupart des systèmes existants de structuration suivent un schéma qui dérive du système de ségrégation de locuteurs proposé par Gish, pour une application liée au trafic aérien et visant à segmenter les dialogues enregistrés entre les contrôleurs et les pilotes en phrases individuelles. Basé sur ce système de ségrégation, l'architecture générale des systèmes de structuration en locuteurs se compose de quatre principaux modules :

- **La paramétrisation du signal** extrait les paramètres acoustiques à partir du signal.
- **La détection de parole** localise les régions de parole dans le flux audio et identifie les régions se composant d'événements acoustiques de non-parole tels que la musique, le

silence ou le bruit.

- **La détection des changements de locuteur** localise les frontières de segments basées sur les changements acoustiques entre différents locuteurs pour assurer que chaque segment contienne de la parole appartenant à un seul locuteur. Cette étape est désigné également sous le nom de la segmentation en locuteur.
- **Le regroupement des segments par locuteur** consiste à regrouper des segments appartenant à un même locuteur

L'architecture présentée ci-dessus est la base d'un système de structuration en locuteurs. Les systèmes existants appliquent cette architecture avec certains modules supplémentaires tels qu'une classification en genre du locuteur (homme/femme) ou du canal de transmission (bande large/étroite) [Gauvain *et al.*, 1998; Tranter *et al.*, 2004] et une étape finale de re-segmentation [Sinha *et al.*, 2005; Meignier *et al.*, 2006]. Les quatre principaux modules présentés dans l'architecture générale ne sont pas tous nécessairement présents dans un système spécifique de structuration. Par exemple, le système de structuration en locuteurs décrit dans [Tranter and Yu, 2003] n'a aucune étape de détection des changements de locuteur et exécute le regroupement des segments directement sur les segments de parole durant moins de 30 secondes avec leurs étiquettes du genre du locuteur et du canal de transmission issues de la classification acoustique. L'ordre des modules peut aussi varier suivant le système. La structure séquentielle entre les modules peut également être remplacée par une forme intégrée de plusieurs modules, tel que le procédé itératif conjoint de segmentation et de regroupement en locuteur présenté dans [Gauvain *et al.*, 1998].

Méthodes d'évaluation et campagnes d'évaluation

La performance du système de structuration en locuteurs est mesurée en suivant le protocole défini dans les évaluations NIST RT. Dans son principe, cette méthode consiste à faire d'abord une correspondance entre les identités des locuteurs données par la transcription de référence et les identités des locuteurs données par le système. La métrique primaire désignée sous le nom d'erreur de locuteur, est la fraction du temps qui n'a pas été attribué par le système à un locuteur correct, étant donnée la correspondance optimale entre locuteurs. Une autre mesure est le taux d'erreur global de structuration en locuteur, i.e. Diarization Error Rate (DER) qui inclut les erreurs de locuteur manqué et de fausse alarme de locuteur, tenant compte de ce fait des erreurs de détection entre parole et non-parole [NIST, 2004].

Les systèmes de structuration en locuteurs développés pendant cette thèse ont été validés au cours de deux séries d'évaluations: les évaluation internationales NIST Rich Transcription (NIST RT) et l'évaluation française ESTER 2005 dans le cadre du projet Technolangue EVALDA.

La série d'évaluations sur la transcription enrichie conduite par le NIST a débuté en 2002, la plus récente à ce jour ayant eu lieu en 2007. L'objectif des évaluations NIST RT est de rendre des transcriptions plus lisibles en intégrant à la transcription orthographique des informations de

méta-données sur le locuteur ou le discours. La tâche de structuration en locuteurs a parfois été présentée dans les évaluations NIST RT comme une sous-tâche de l'extraction de méta-données (MDE). Nous avons participé à la structuration en locuteurs dans l'évaluation de NIST Rich Transcription 2004 Fall (RT-04F) sur des enregistrements de journaux télévisés en anglais, ainsi que dans les évaluations NIST RT sur la transcription enrichie de réunion en 2006 et 2007, désignées respectivement sous les nom de "RT-06S" et de "RT-07S", sur des enregistrements de réunions en anglais.

La campagne d'évaluation ESTER a été organisée dans le cadre du projet EVALDA consacré à l'évaluation des technologies de la parole pour la langue française, sous l'égide scientifique de l'Association Francophone de la Communication Parlée (AFCP) avec le concours du Centre d'Expertise Parisien de la Délégation Générale de l'Armement (DGA/CEP) et de ELDA (Evaluations and Language resources Distribution Agency). Cette campagne a débuté en 2003 et s'est composée d'une phase I de test à blanc en janvier 2004 et d'une phase II d'évaluation officielle en janvier 2005. Les données utilisés dans ESTER ont été tirées de différentes émissions radiophoniques en francais.

Structuration en locuteurs pour les émissions d'actualités

Le système de structuration en locuteurs pour les émissions d'actualités radio et télé-diffusées (Broadcast News ou BN) développé pendant cette thèse se fonde sur un système de division de l'audio initialement conçu comme une étape de prétraitement pour le système de transcription automatique du LIMSI [Gauvain *et al.*, 1998; Gauvain *et al.*, 2001]. Ce système de base fournit des classes de locuteur assez pures, avec une tendance à diviser les segments appartenant aux locuteurs ayant une grande quantité de données en plusieurs classes. Plusieurs améliorations ont été appliquées au système de base. D'abord, la procédure itérative par mélange de Gaussiennes (GMMs) maximisant conjointement la segmentation et le regroupement a été remplacée par une classification hiérarchique par agglomération basée sur le Critère d'Information Bayesien (BIC). En second lieu, une deuxième étape a été intégrée afin de regrouper les classes issues du regroupement BIC en utilisant des méthodes à l'état de l'art en identification du locuteur (SID). Enfin une étape de post-traitement raffine les frontières de segment à partir du résultat d'un système de transcription.

Le système de structuration en locuteurs présenté traite des vecteurs acoustiques de 38 dimensions qui se composent de 12 coefficients cepstraux, de leurs coefficients différenciels Δ and Δ-Δ plus des Δ et Δ-Δ de la log-énergie. Le système primaire est structuré comme suit:

- **Détection de la parole**: la parole est extraite à partir du signal par un décodage de Viterbi en utilisant 5 GMMs pour les catégories de parole, parole avec bruit, musique avec parole, musique seule, et silence ou bruit. Chacun de ces GMMs se compose de 64 Gaussiennes entraînées sur environ 1 heure de données, choisie parmi des enregistrements de journaux télévisés distribués par le Linguistic Data Consortium (LDC).

- **Détection des changements acoustiques**: il est effectuée sur la totalité du signal afin de trouver les points de changements entre toutes les sources audio possibles (par exemple, locuteur, musique, silence et bruit de fond), ainsi que les changements d'environnement sonore au cours d'un même tour de parole. Ceci est fait en prenant les maxima d'une mesure locale de divergence gaussienne entre deux fenêtres adjacentes de 5 secondes qui sont appliquées par fenêtre glissante sur le signal entier. C'est une méthode similaire à la segmentation basée sur la métrique KL2 (Kullback Leibler symétrisée) [Siegler *et al.*, 1997]. Chaque fenêtre est modélisée par une Gaussienne diagonale en utilisant les coefficients statiques (c.-à-d. seulement les 12 coefficients cepstraux plus l'énergie).

- **Resegmentation par Viterbi**: chaque segment issu de la détection des changements acoustiques est modélisé par un GMM à 8 Gaussiennes avec matrice de covariance diagonale. Ensuite, les frontières des segments de parole détectés par le module de détection de parole sont raffinées en utilisant un décodage de Viterbi avec cet ensemble de GMMs.

- **Étiquetage en genre du locuteur et en canal de transmission**: la détection du canal de transmission (bande large ou étroite) pour chaque segment est d'abord effectuée en utilisant la classification par le maximum de vraisemblance avec deux GMMs indépendants du locuteur. L'identification du genre du locuteur (homme ou femme) est alors effectuée sur les segments en utilisant deux paires de GMMs dépendant de la largeur de bande. Les GMMs se composent de 64 Gaussiennes avec matrice de covariance diagonale et ont été entraînés sur le même ensemble de données que celui qui a été employé pour entraîner les modèles de détection de la parole.

- **Regroupement BIC**: un groupe initial de locuteurs c_i est modélisé par une Gaussienne avec une matrice de covariance pleine Σ_i estimée sur les trames acoustiques de chaque segment n_i issu de la resegmentation par Viterbi. Le critère BIC [Chen and Gopalakrishnan, 1998a] est employé comme mesure de distance inter-classes et comme critère d'arrêt. Le critère BIC est défini comme:

$$\Delta BIC = (n_i + n_j)\log|\Sigma| - n_i \log|\Sigma_i| - n_j \log|\Sigma_j| - \lambda\frac{1}{2}(d + \frac{1}{2}d(d+1))\log n \quad (1)$$

  où d est la dimension des vecteurs de paramètre, $n = n_i + n_j$ et λ le poids affecté à la pénalité. À chaque itération, les deux groupes les plus proches (c.-à-d. ceux dont la valeur de ΔBIC est la plus négative) sont agglomérés et les valeurs entre le nouveau groupe et les groupes restants sont recalculées. Ce procédé de regroupement s'arrête quand tout ΔBIC est supérieur à zéro.

- **Regroupement SID**: après l'étape du regroupement BIC, des méthodes d'identification du locuteur [Schroeder and Campbell, 2000; Barras and Gauvain, 2003] sont employées pour améliorer la qualité du regroupement de locuteurs. La normalisation des vecteurs acoustiques ("feature warping" [Pelecanos and Sridharan, 2001]) est effectuée sur chaque segment en utilisant une fenêtre glissante de 3 secondes afin de rendre la distribution

des coefficients cepstraux identique à une distribution normale et réduire les effets non-stationnaires de l'environnement acoustique. Le GMM de chaque groupe est obtenu par l'adaptation du maximum a posteriori (MAP) [Gauvain and Lee, 1994] des moyennes du modèle de référence UBM (Universal Background Model [Reynolds *et al.*, 2000]) correspondant. Pour chaque combinaison de genre du locuteur et de canal de transmission, un UBM à 128 Gaussiennes diagonales est entraîné sur plusieurs heures de données de journaux télévisés ou radiophoniques en anglais diffusées entre 1996 et 1997. Une deuxième étape de regroupement agglomératif est alors exécutée en utilisant le rapport de log-vraisemblance croisé entre deux classes c_i et c_j similaire à celui présenté dans [Reynolds *et al.*, 1998]:

$$\mathcal{S}(c_i, c_j) = \frac{1}{n_i} log \frac{f(x_i|M_j)}{f(x_i|B)} + \frac{1}{n_j} log \frac{f(x_j|M_i)}{f(x_j|B)} \quad (2)$$

où $f(x_k|M)$ est la probabilité des trames acoustiques de la classe c_k sachant le modèle M, et n_k est le nombre de trames dans la classe c_k. Le regroupement en locuteurs s'arrête quand le rapport de log-vraisemblance croisé entre toutes les classes est inférieur à un seuil δ donné.

- **Post-filtrage des silences**: la segmentation en mots générée par le système de transcription du LIMSI [Nguyen *et al.*, 2004] est employée dans une étape de post-traitement pour filtrer les segments de silence de courte durée qui n'ont pas été détectés par l'étape de détection de la parole. Seuls le silence entre mots plus longs qu'une seconde sont exclus, cette valeur ayant été déterminée empiriquement.

Le système amélioré de structuration en locuteurs a été employé pour participer à l'évaluation NIST RT-04F ainsi qu'à l'évaluation ESTER. Bien que les enregistrements de journaux télévisés ou radiophoniques utilisés dans les deux évaluations soient en différentes langues (c.-à-d. l'anglais américain pour RT-04F et le français pour ESTER), le système développé présente des performances au niveau de l'état de l'art et a obtenu les meilleurs résultats dans les deux évaluations. Les expériences faites sur les données de développement de RT-04F prouvent qu'une réduction relative de 70% du taux d'erreur est réalisée par le système amélioré comparé au système de base. Les résultats obtenus sur les données de développement de RT-04F et d'ESTER montrent que le regroupement SID peut conduire à une réduction d'erreur de manière significative comparé au système utilisant seulement l'étape du regroupement BIC. Sur les données de test des deux évaluations, le système amélioré présente des taux d'erreurs similaires de 9% environ lorsque les paramètres du système ont été optimisés pour les données. Dans l'évaluation RT-04F, une réduction supplémentaire d'erreur de 0.6% absolu a été obtenue en enlevant les silences entre les mots générés par le système de transcription automatique du LIMSI.

Les performances du système de structuration développé dépendent fortement des valeurs des paramètres du système (c.-à-d. le poids de la pénalité λ pour le critère BIC et le seuil δ existant dans le regroupement SID) qu'il est nécessaire d'adapter sur des données spécifiques. Une étude préliminaire sur la robustesse du système est présentée dans cette thèse, en se concentrant sur la

corrélation entre les valeurs optimales des paramètres du système et la durée des enregistrements. Les données examinées ont été extraites des données d'apprentisage d'ESTER et se composent de 20 fichiers audio d'une heure diffusés sur "France inter". Des expériences contrastives sont effectuées sur le sous-ensemble comprenant les premières 30 minutes et les dernières 30 minutes d'audio. La durée de l'enregistrement n'apparaît pas comme un facteur important qui influence considérablement les valeurs optimales des paramètres du système, bien que le seuil du regroupement SID tende à être plus petit sur des émissions de courte durée. Les résultats obtenus en utilisant seulement le regroupement BIC montrent que la pénalité de BIC devrait être modifiée pour prendre en compte des facteurs liés aux caractéristiques des enregistrements, en plus de la dimensionnalité des vecteurs acoustiques et de la taille de données.

Des émissions d'actualités aux réunions

Beaucoup d'efforts de recherches sur la structuration en locuteurs se sont d'abord concentrés sur le domaine des journaux télévisés ou radiophoniques, aussi de nombreuses méthodes ont été proposées et on montré leur efficacité dans ce domaine [Tranter *et al.*, 2004; Barras *et al.*, 2006]. Ces dernières années, l'intérêt des recherches a évolué vers les enregistrements de réunions [Anguera *et al.*, 2006c; Wooters and Huijbregts, 2007], qui posent plus de difficultés pour la structuration en locuteurs. La parole au cours des réunions est complètement spontanée, avec des périodes fréquentes de parole superposée et un grand nombre de pauses pour tous les locuteurs. En outre, les réunions sont souvent enregistrées en utilisant différents types de microphones situés à diverses positions dans la salle, fournissant de multiples fichiers audio aux qualités acoustiques différentes pour une même réunion. Enfin, l'utilisation des microphones éloignés rend également le signal acoustique des réunions plus bruyant que la plupart des enregistrements de journaux télévisés.

Dans le cadre des évaluations sur la transcription enrichie de réunions organisées par le NIST [NIST, 2006; NIST, 2007], les enregistrements de réunions ont été divisés en deux sous-domaines: les conférences et les séminaires. Comparé aux enregistrements de conférence, les séminaires contiennent moins d'interaction entre les participants, et se composent typiquement d'un exposé par un présentateur suivi d'une session de question/réponse ou de discussion.

Il y a différents types de microphones utilisés dans chaque salle de réunion, et les enregistrements de chaque réunion sont fournis sous la forme d'un ensemble de signaux synchronisés. Dans l'évaluation NIST RT-06S et RT-07S, plusieurs configurations de prise de son sont proposées pour la tâche de structuration en locuteurs. Nos systèmes de structuration proposés pour les réunions ont été validés sur les conditions de prise de son distante dans ces deux évaluations: microphone éloigné unique (SDM) et microphones éloignés multiples (MDM). En outre, nous avons également participé à l'évaluation RT-06S pour la prise de son par réseau multiple de microphones Mark III (MM3A), sur la sortie du traitement d'antenne fourni par l'université de Karlsruhe sur les 64 canaux Mark III.

Comme le système de structuration en locuteurs développé pour les journaux télévisés est con-

struit en utilisant des modules séparés, il est possible de garder la même structure de système et de faire les adaptations nécessaires aux modules répondant aux nouvelles exigences du domaine des réunions. Puisque les données de réunion incluent souvent des segments de parole de courte durée, le système amélioré de structuration pour journaux télévisés présente un taux erreur élevé en détection de parole, en particulier beaucoup d'erreurs de parole manquée. Ce fait dérive de l'absence de contrôle temporel pour chaque modèle dans le décodage standard de Viterbi. Une pénalité de transition peut être employée pour garantir la durée des segments, mais avec l'augmentation du niveau de bruit, la probabilité du modèle de parole diminuera et les segments de parole les plus courts seront ainsi rejetés. Un détecteur de parole plus simple basé sur le rapport de log-vraisemblance (LLR) a alors été choisi. Ce détecteur calcule le LLR entre les modèles de parole et non-parole pour chaque trame acoustique et remplace le décodage de Viterbi par une comparaison de LLRs lissés. Différentes probabilités a priori peuvent être données aux modèles de parole et de non-parole lors de l'application du détecteur fondé sur le LLR.

Les expériences faites sur les séminaires des données de développement de RT-06S prouvent que le détecteur LLR a fourni une réduction significative d'erreur, jusqu'à 58% en relatif par rapport à la détecion de parole par Viterbi. Sur les séminaires provenant des données d'évaluation de RT-06S, le système de structuration intégrant le détecteur par LLR a permi un DER de 24,5% sur la condition SDM, un résultat similaire à celui obtenu sur la condition MM3A. Un meilleur résultat de 21,5% DER a été obtenu sur la condition MDM. Une raison possible de cette diminution de l'erreur est que le signal choisi aléatoirement pour MDM a une meilleure qualité que celui de la condition SDM.

D'autres améliorations ont été réalisées au cours de cette thèse pour développer un système commun de structuration pour les réunions de type conférence et de type séminaire. Puisque le système de structuration développé pour l'évaluation RT-06S fournit une erreur de détection de parole toujours élevée, différentes représentations acoustiques avec diverses techniques de normalisation ont été étudiées dans le cadre de la détection de parole avec le LLR. Nous avons proposé une nouvelle méthode de normalisation d'énergie tenant compte du facteur de voisement, bien que les expériences sur les données de développement de RT-07S prouvent que l'algorithme de normalisation proposé marche moins bien que la technique de normalisation par la variance. Des modèles UBMs entraînés sur différents types de représentations acoustiques ont été également examinés lors de l'étape de regroupement SID. Les modèles de parole/non-parole utilisés dans la détection de la parole et les UBMs utilisé à l'étape de regroupement SID ont été entraînés sur une sélection de données de réunions utilisées dans les précédentes évaluations RT, avec les transcriptions alignées correspondantes des transcriptions de référence. Le système de structuration pour réunions a utilisé également un traitement d'antenne des signaux produit par le système de delay&sum [Anguera *et al.*, 2005a] développé chez ICSI (International Computers Science Institue) pour la condition d'entrée audio MDM.

Le détecteur de parole par LLR utilisant des modèles de parole et de non-parole entraînés sur des paramètres acoustiques normalisés par la variance a donné les meilleurs résultats en détection de parole sur les données de développement de RT-07S, c.-à-d. 3,9% pour les enregistrements

de conférences et 6,6% pour les enregistrements de séminaires. La disparité entre les données d'apprentissage de conférence et les données de test de séminaire conduit à une erreur de détection de la parole relativement plus élevée sur les enregistrements de séminaires. Quant aux modèles UBM, le meilleur résultat a été produit en employant un modèle indépendant du genre du locuteur estimés sur 12 coefficients cepstraux en échelle de fréquence Mel plus les coefficients différentiels Δ et le Δ de la log-énergie avec la technique de normalisation "feature warping". Lors des évaluations RT-07S, le système de structuration adapté aux réunions a fourni des résultats comparables sur la sortie du traitement d'antenne en condition MDM pour les conférence et pour les séminaires (c.-à-d. un DER de 26,1% pour les conférences et de 25,8% pour les séminaires). Le taux d'erreur a augmenté à 29,5% pour les conférences en condition SDM, alors que pour séminaires en condition SDM, le taux d'erreur reste très proche de celui obtenu pour la condition MDM.

Détection de la parole superposée

La parole superposée est souvent ignorée par les systèmes de structuration en locuteurs. Ce peut être un choix acceptable pour des journaux télévisés où les conditions d'enregistrement sont controlées, mais dans des situations où la parole est plus spontanée comme dans les enregistrements de réunions, la proportion de parole superposée augmente considérablement. Une recherche préliminaire sur la détection de parole superposée a été effectuée pendant le travail de cette thèse pour un cas relativement simple, c.-à-d. de la parole téléphonique conversationnelle (CTS). Les données CTS ont été choisies car il y a généralement deux locuteurs présents dans une conversation téléphonique et nous disposons de bases de données dans lesquelles chaque locuteur est enregistré dans un canal séparé. Ceci permet de créer un modèle de parole superposée obtenu en mélangeant les signaux des deux locuteurs. Il sera possible d'adapter l'algorithme proposé pour la détection de la parole superposée au cas des enregistrements de réunions s'il fournit un performance raisonnable de détection sur les données CTS.

La méthode proposée pour la détection de parole superposée est fondée sur l'hypothèse que les données d'apprentissage pour deux locuteurs dans une conversation téléphonique sont disponibles ainsi que la segmentation de référence correspondante. Ces informations a priori peuvent être acquises en utilisant la première moitié d'une conversation comme données d'apprentissage et le reste comme données de test. Étant donnés deux enregistrements avec canal séparé d'une conversation téléphonique, un signal mélangé est obtenu par addition des deux canaux. Deux modèles respectivement pour la parole non-superposée et la parole superposée sont d'abord entraînés sur la première moitié de la conversation. Le modèle de parole non-superposée est entraîné sur les segments de voix provenant de chaque locuteur extrait à partir du signal mélangé et le modèle de la parole superposée est estimé sur les données qui sont obtenues en mélangeant seulement les échantillons de la parole de chaque locuteur extrait à partir des différents canaux. La détection de parole superposée est alors effectuée en calculant le rapport de vraisemblance entre les modèles de parole non-superposée et de parole superposée sur la deuxième moitié de la conversation.

Les expériences ont été faites sur un sous-ensemble de 50 appels extrait du corpus d'enregistrements à partir de téléphone cellulaire distribué par le LDC. Dans le cadre de la détection de parole superposée fondée sur le rapport LLR, nous avons étudié des coefficients cepstraux en échelle de fréquence Mel (MFCC) et des paramètres d'autocorrélation aussi bien que leur combinaison. Les premiers résultats expérimentaux montrent que la méthode proposée avec les paramètres d'autocorrélation a une meilleure performance que celle avec des coefficients cepstraux en échelle Mel. Le système combiné montre un résultat de détection similaire à celui obtenu avec les seuls paramètres d'autocorrélation, et ces deux systèmes donnent le taux d'erreur égal minimum (EER) de 33,6% pour la décision de parole superposée contre parole non-superposée. La détection de parole superposée semble plus facile dans les conversations entre locuteurs féminins que lorsque les locuteurs sont de genres différents.

Perspectives

Les recherches actuelles sur la structuration en locuteurs ont atteint de bonnes performances pour des enregistrements de journaux télévisés, alors que les résultats sur des enregistrements de réunions sont moins bons. Il y a donc un grand potentiel d'amélioration des performances pour notre système de structuration de réunions. Comme des signaux multiples sont normalement disponibles pour une réunion, il est possible d'exploiter l'information multicanale pour aider la structuration en locuteur. Par exemple, des modèles entraînés sur les différences de temps d'arrivée entre les différents canaux peuvent être combinés avec les modèles acoustiques classiques dans le cadre d'un regroupement BIC [Pardo *et al.*, 2006a; Pardo *et al.*, 2006b]. Une autre caractéristique des données de réunions est de contenir un grand nombre de zones de silence de courte durée. Les modèles des locuteurs peuvent être corrompus par ces silences, de ce fait dégradant la performance du regroupement de locuteurs. Un algorithme de purification visant à supprimer les segments de non-parole existantes dans les classes à regrouper peut être utile pour le regroupement de locuteurs[Anguera *et al.*, 2006b].

Un autre problème restant dans le domaine de structuration en locuteur est la détection de la parole superposée. À partir de l'évaluation NIST RT-06S, la métrique primaire a été calculée sur la totalité du document y compris les parties superposées. On exige ainsi des systèmes de structuration qu'ils détectent les segments où plusieurs locuteurs parlent en même temps et qu'ils produisent l'identité correspondant à chaque locuteur. Plusieurs chercheurs ont proposé des méthodes pour détecter la parole superposée y compris l'auteur de cette thèse, mais jusqu'à présent aucun gain significatif n'a été atteint. L'information sur les positions de locuteurs extraites à partir des prises de sons multiples peut être utile pour la détection de la parole superposée.

A partir des expériences présentées dans cette thèse, il apparaît que les paramètres existant dans les systèmes de structuration sont sensibles aux données traitées et doivent être ajustés avec précision sur des données de développement. La disparité entre les données de développement et de test influence la performance de la structuration. L'amélioration de la robustesse des systèmes est une direction de recherches d'intérêt à l'avenir. Le travail potentiel dans cette direction est de

faire des systèmes de structuration choisissant des valeurs optimales des paramètres automatiquement à partir des données. Pour ce faire, d'autres investigations sont nécessaires pour découvrir la corrélation fondamentale entre les paramètres du système et les caractéristiques acoustiques des enregistrements audio. L'autre approche possible est de réduire le nombre de paramètres de système ou la quantité de données d'apprentissage nécessaire pour l'entraînement des modèles acoustiques.

Part II

Acoustic-based Speaker Diarization

Chapter 1

Introduction

With the continually decreasing cost of storage capacity and the developments in processing power, the quantity of digital information increases rapidly in the recent decades. These digital information may be stored in different types of media such as text, picture, audio and video and so on. Efficient and effective content based access to information becomes an important task.

This thesis focuses on audio recordings including broadcasts news of radio or television, telephone and meetings. The various indexing algorithms facilitating the retrieval of relevant information provide easy access to audio documents. The conventional audio indexing technique consists of running an automatic speech recognizer (ASR) to extract the words spoken in the audio and then applying some text-based information retrieval approaches to the output transcription. However, this kind of transcripts is often difficult to understand and do not include all the information contained in the spoken document. The Rich Transcription (RT) evaluations [NISTRT] organized by NIST (National Institute for Standards and Technology) propose to add meta-data information into the simple word sequence and make the transcripts more readable. The first group of meta-data is the acoustic based information such as speaker turn, the number of speakers, speaker gender, speech bandwidth and other sounds (e.g. music, noise) etc. The second group of meta-data is the information related to the nature of spontaneous speech like disfluencies (e.g. hesitations, repetition and revision) and emotions. The third type of meta-data is the linguist information such as named entity and topic extraction.

The topic of this thesis is the extraction of the meta-data related to speaker identity by using only acoustic characteristics of the audio. Although one can also take advantage of linguistic information for the extraction of true speaker identity [Canseco-Rodriguez *et al.*, 2004; Canseco-Rodriguez *et al.*, 2005; Canseco-Rodriguez, 2006], it falls outside of the scope of this thesis. This task aims to answer the question of "who spoken when" given an audio document and is referred to as "speaker diarization" in the NIST Rich Transcription evaluations. The speaker diarization is the process of partitioning an audio input stream into homogeneous segments according to speaker identity. The result of a speaker diarization system is a sequence of speech segments which boundaries are located at speaker change or speech/non-speech change points, with their

associated cluster labels. Each segment cluster is assumed to represent one speaker occurring in the audio, whatever the amount of data each speaker has.

The speaker diarization is one aspect of audio diarization that is defined as the process of annotating an input audio stream with information that attributes temporal regions of signal to their specific sources [Tranter and Reynolds, 2006]. These sources can include particular speakers, music, background noise sources and other source characteristics. The types of the audio sources are application specific that could be defined as very broad or concrete acoustic classes. The simplest case of audio diarization is the speech and non-speech detection.

Up to now, there are three primary domains that are used for speaker diarization research:

- Conversational telephone speech (CTS): involving two speakers and using single channel speech signals, the speaker diarization was first evaluated within the NIST speaker recognition evaluation from 2000 to 2002 [NISTSRE]
- Broadcast News (BN) from radio and TV: including more complicate acoustic natures such as studio/telephone speech, music, speech over music etc. and multiple speakers in shows. This data type became the main research domain for speaker diarization from 2002 to 2004 in the NIST Rich Transcription evaluations [NIST, 2003; NIST, 2004]
- Meetings: consisting of lectures and conferences data sub-types, where the speech is completely spontaneous with frequent periods of overlapping and the signals present several distortions and noisy environment due to the use of distant microphones. This domain has received more research attention in the NIST Rich Transcription meeting recognition evaluations from 2004 to 2007 [NIST, 2005; NIST, 2006; NIST, 2007], with the cooperation of the European projects CHIL (Computers in the Human Interaction Loop) [1] and AMI (Augmented Multiparty Interaction)

The three domains described above present different challenges for the speaker diarization task, as they differs in the number of speakers, signal quality, types of acoustic sources, the durations of speaker turns and speech style. In general, speaker diarization systems are designed towards specific domain. More recently, some research work have been done to develop more portable diarization system across different domains [Anguera *et al.*, 2006c; Wooters and Huijbregts, 2007].

The speaker diarization is also called as speaker segmentation and clustering in some literatures, as most diarization systems rely on a two-steps structure in which the speaker segmentation is followed by a clustering process. The speaker segmentation step is also referred to as speaker change detection and aims to locate the boundaries of speech segments by finding the speaker change or more generally acoustic change points. The objective of the acoustic change detection is to find the change points between all possible audio sources (e.g. speaker, music, silence and

[1] The work of this thesis is totally financed by the European Commission under the FP6 Integrated Project IP 506909 CHIL

background noise), besides the changes between speakers. Furthermore, the acoustic change detection also searches the changes of the background environment in a single speaker turn. The speaker clustering is then applied to the speech segments issued from the segmentation stage for grouping the segments coming from the same speaker.

The speaker diarization is a very useful processing for various speech technologies as follows:

- Indexing of audio documents: speaker diarization is of interest for the indexing of the audio according to the speakers, which facilitates the processing, searching, accessing of the content related to a particular speaker.
- Rich transcription: speaker diarization can be used to improve the readability of an automatic transcription by structuring the audio stream into speaker turns and providing speaker identity that is a cluster label in most cases and may be the true identity in some particular cases. This rich transcription is potentially helpful for other speech technologies such as summarization, translation and parsing.
- Automatic speech recognition: on the one hand, segmenting the audio into small segments and separating speech/non-speech segment is helpful to reduce the computation time of the recognition process and avoid the word insertions in these portions; on the other hand, clustering segments from the same speaker increases the amount of data available for speaker adaptation, so that the recognition performance can be improved significantly.
- Speaker recognition: speaker diarization provides speech segments containing the voice of only one speaker, which is very useful for some speaker related tasks such as speaker identification, speaker verification and speaker tracking.

The difficulties for the speaker diarization task depend largely on the amount of allowed prior knowledge. Various prior knowledge can be available in some specific diarization applications. These may be sample speech from each speaker (e.g. meeting participants in regular group meetings) or some speakers (e.g. famous anchors on particular TV or radio stations) in the audio. In this case, the speaker diarization task turns into a speaker tracking task [Bonastre *et al.*, 2000; Istrate *et al.*, 2005b]. The prior knowledge could also be the number of speakers in the audio, for example, two speakers involved in conversational telephone speech and the structure of audio such as a presentation followed by a discussion/question session in lecture recordings. However, the work presented in this thesis is based on the definition proposed by NIST Rich Transcription evaluations, where no *a priori* knowledge of the speakers voice or even of the number of the speakers is available. By following this assumption, the diarization system produces only a relative, show-internal speaker identification.

The manuscript of this thesis is composed of 6 chapters and the remainders of it are outlined as follows:

Chapter 2 reviews the state of the art in the speaker diarization domain. A general architecture of speaker diarization system is introduced at the beginning of this chapter. Some fundamental

speech processing techniques used for the speaker diarization are given in the subsequent part. The emphasis of this chapter is to provide an overview of the commonly used approaches for the key modules in the general diarization framework as well as their implementations in current systems. Chapter 3 presents the metrics used to evaluate the diarization performance and the histories of the speaker diarization evaluations within the international NIST Rich Transcription and the French Technolangue ESTER evaluation campaigns.

Chapter 4 first describes the baseline diarization system that is designed as a preprocessing for Broadcast News transcription system. This baseline system relies on an iterative segmentation/clustering procedure and tends to split data from the same speaker into small segments according to the change of background environment. Then, the improvements to the baseline BN diarization system introduced in this thesis are detailed. The improved multi-stage speaker diarization system combines a Bayesian Information Criterion (BIC) agglomerative clustering and a second clustering stage relying on a GMM-based speaker identification (SID) method. This improved BN speaker diarization system has given the best diarization performance in the NIST Rich Transcription (RT) 2004 Fall diarization evaluation on US English test data and also in the ESTER evaluation on the French test data. The results obtained from the baseline and improved diarization systems are also compared in this chapter. Finally, a preliminary study of the system robustness is presented by investigating the correlation between system parameters and the duration of BN shows.

Chapter 5 presents the work that is done for adapting the multi-stage diarization system from Broadcast News to meetings. In order to address this problem, the analysis of the differences between BN and meetings is given at the beginning of this chapter. The developments to the baseline BN diarization system for the lecture data used in the NIST RT 2006 Spring (RT-06S) meeting recognition evaluation are then presented. These include a new speech activity detection based on log-likelihood ratio. The further work intended to develop a common speaker diarization system for both conference and lecture data is also described in this chapter. These focus on the exploitation of various set of features in the speech activity detection and the SID clustering. The experimental results obtained on the data of both RT-06S and RT-07S evaluation are shown in the end of this chapter.

Chapter 6 aims to the frequent portions of overlapping speech occurring in meeting recordings and describes a preliminary overlapping detection method for conversational telephone speech, since it is assumed there are only two speakers in CTS data. The proposed detection algorithm is based on the use of speech sample from each speaker and training of the overlapping speech model from separate channel recordings. The overlapping speech detection is carried out by computing log-likelihood ratio between single speaker and overlapping speech models. Different sets of features are investigated within this experiment framework.

Finally, the contributions of this work and the primary observations are summarized in Chapter 7. Some conclusions and prospections in the future work are also given.

Chapter 2

Background

An overview of the automatic speaker diarization is provided in this chapter. A general architecture of speaker diarization systems is presented at the beginning of this chapter. The fundamental speech processing techniques widely used in speaker diarization are subsequently described: speech parameterization and statistical speaker modeling methods such as the Gaussian mixture model. The most part of this chapter is focused on introducing the common approaches employed in the principal modules of the general diarization architecture, as well as their different implementations in current speaker diarization systems.

2.1 General architecture of speaker diarization systems

Speaker diarization, also called speaker segmentation and clustering, is the process of partitioning an input audio stream into homogeneous segments according to speaker identity. It is one aspect of audio diarization, along with categorization of music, background noise and channel conditions. The aim of the speaker diarization is to provide a set of speech segments, where each segment is bounded by speaker change or speech/non-speech change points and labeled with the identity of the speaker engaging in the corresponding speech.

The two main applications of speaker diarization can be found in the literature. On the one hand, speaker diarization is a very useful preprocessing step for Automatic Speech Recognition (ASR) systems. By separating out speech and non-speech segments, the recognizer only needs to process audio segments containing speech, thus reducing the computation time. By clustering segments of the same acoustic nature, condition specific models can be used to improve the quality of the recognition. By clustering segments from the same speaker, the amount of data available for unsupervised speaker adaptation is increased, which can significantly improve the recognition performance. On the other hand, speaker diarization is beneficial to rich transcription and speaker indexing systems. The speaker diarization techniques can improve readability of an automatic transcription by structuring the audio stream into speaker turns and in some cases by

providing the identity of the speakers. Such information can also be of interest for the indexing of multimedia documents.

The currently used speaker diarization techniques are largely dependent on the statistical and probabilistic methods used for modeling the differences of the acoustic characteristics.

Most current speaker diarization systems are founded on a framework derived from the speaker segregation system proposed by Gish [Gish *et al.*, 1991]. This segregation system was developed by the *BBN* company for an air traffic control application and aimed to divide the recorded dialogs between controllers and pilots into the speech utterances from the individuals. Because an air traffic controller could use the same frequency to speak with different pilots, it was possible that a controller spoke to several pilots in one conversation recording. The purpose of this application was to automatically recognize the commands from the controller to each pilot. Hence, a speaker segregation system was needed before the speaker identification and speech recognition processing.

Based on the segregation system proposed by Gish, the general architecture of current speaker diarization systems is composed of four principal modules (c.f. Figure 2.1):

- **Front-end parameterization** extracts the acoustic parameters from the speech signal.
- **Speech activity detection** locates the regions of speech in the input audio stream and removes the regions consisting of non-speech acoustic events such as music, silence or noise.
- **Speaker change detection** locates the segment boundaries based on the acoustic changes between different speakers so that each segment consists of the speech derived from a single speaker. This step is also referred to as speaker segmentation.
- **Speaker clustering** regroups the segments coming from the same speaker into a cluster.

The architecture illustrated in Figure 2.1 is a basis of speaker diarization, the current diarization systems are developed by completing this architecture with some supplementary modules. For example, some researchers improve the basic architecture by incorporating a gender/bandwidth classification [Gauvain *et al.*, 1998; Tranter *et al.*, 2004; Barras *et al.*, 2006], as well as employing a re-segmentation as the final step of diarization systems [Sinha *et al.*, 2005; Zhu *et al.*, 2005; Meignier *et al.*, 2006]. The four principal modules presented in the general architecture are not all required to occur in a specific diarization system. For example, the speaker diarization system described in [Tranter and Yu, 2003] has no speaker change detection stage and performs the speaker clustering directly on the speech segments lasting less than 30 seconds with gender and bandwidth labels produced by the acoustic classification. The ordering of the modules can be varied in different speaker diarization systems. The sequential structure between modules can also be replaced by an integrated form of different modules, such as the iterative speaker segmentation and clustering procedure presented in [Gauvain *et al.*, 2001]. The different implementations of the speaker diarization systems will be explained later in this chapter.

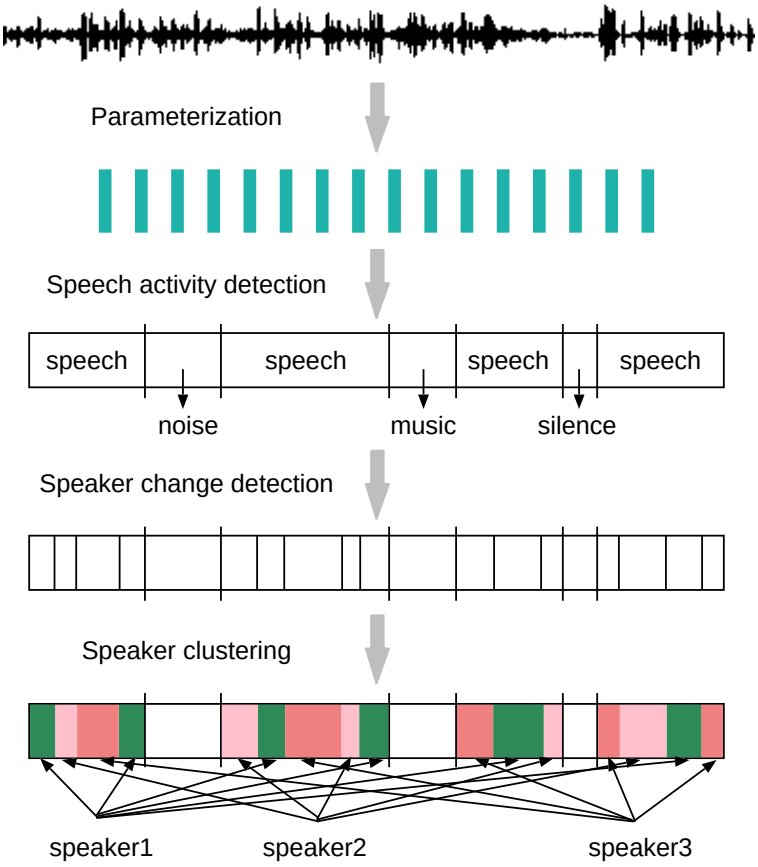


Figure 2.1: An illustration of the general architecture of speaker diarization systems.

Although prior knowledge can be allowed in some specific speaker diarization applications (for example, the prior information of the familiar anchors voice on particular news stations), this thesis will adopt the following definition: no *a priori* knowledge of the speakers voice or even of the number of the speakers is available. This is the diarization task definition used in the Rich Transcription diarization evaluations since 2000 [NIST, 2000]. Under this assumption and considering that the speaker diarization task is relative to a given audio recording, only the relative, recording-internal speaker identities are produced by the diarization system.

2.2 Grounds for speaker diarization

In this section, we introduce some fundamental speech processing techniques which can be useful for speaker diarization. The parameterization of speech signals and the statistical speaker modeling techniques will be described respectively.

2.2.1 Front-end parameterization

As mentioned previously, a fundamental module of speaker diarization system is the front-end parameterization which converts the speech signal into a set of feature vectors. The speech signal could not be directly used by a diarization system because of both the large variety of sound information existing in the signal and the limitation of the mathematical methods to deal with this variety. Therefore a more compact and efficient representation of the speech signal is necessary. The goal of speech parameterization is to extract sound features from the speech signal relevant for the problem to be solved.

The speech signal is typically a non-stationary signal, since it varies over the time reflecting the speech sound being spoken. However, because the production process of speech is closely related to the motion of the human vocal apparatus, and this physical movement is much slower than the vibration of sound pressure, the speech signal can be approximately considered as stationary (i.e. the signal characteristic is relatively constant) over a sufficiently short period of time (between 5 to 100 milliseconds) [Rabiner and Juang, 1993]. Most analysis methods of the speech signal are based on this short-term stationary assumption.

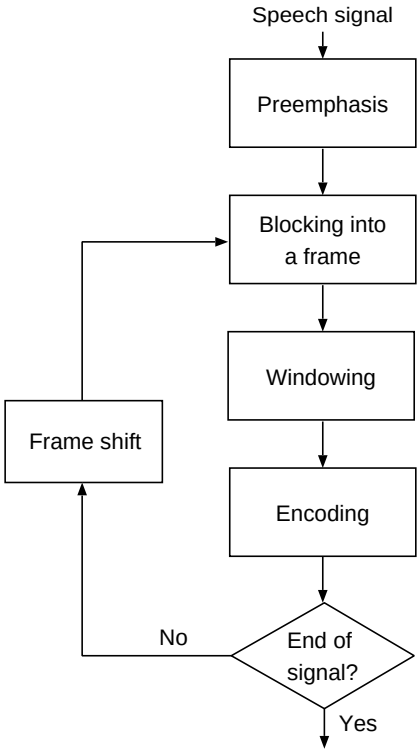


Figure 2.2: Block diagram of a general front-end parameterization processor.

Figure 2.2 shows a block diagram of a general front-end parameterization processor. The speech signal is first preemphasized by performing high frequencies amplification. This preemphasis

aims to compensate the attenuation of the voiced sections in speech signal, which is caused by physiological characteristics of the speech production system [Markel and Gray, 1976]. The preemphasis $s_p(n)$ of a digitized speech signal $s(n)$ is computed as:

$$s_p(n) = s(n) - \alpha s(n-1) \tag{2.1}$$

where α is generally set to between 0.9 and 1.0.

Since the speech signal is non-stationary, the analysis must be performed on short time segments. Therefore the preemphasized speech signal is blocked into frames by the application of a sliding window. The frame duration typically ranges between 20 ms and 50 ms, over which the signal is quasi-stationary according to the short time stationary assumption given above. However the blocking of the speech signal produces a discontinuity of the frame boundaries. In order to reduce this effect, a frame of the speech signal is weighted by a window to tap the signal to zero at the boundaries. Let $w(n)$ be a window consisting of N_w signal samples, the resulting signal $s_w(n)$ from this windowing is defined by the following equation:

$$s_w(n) = s_p(n)w(n), \quad 0 \leqslant n < N_w \tag{2.2}$$

One of the most used windows is the Hamming window [Harris, 1978]:

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N-1}\right), \quad 0 \leqslant n < N_w \tag{2.3}$$

Then the windowed signal frame is encoded into a set of parameters. After the sliding window is shifted in a typical step of 10 ms, the same chain of operations is repeated until the end of the speech signal is attained.

There are two main analysis approaches to encode the speech signal: non-parametric representation and parametric modeling. The speech signal can be represented mathematically on the temporal or spectral domain without modeling it. The most typically non-parametric representations are the short time Fourier transform [Oppenheim and Schafer, 1975; Rabiner and Schafer, 1978] and digital filter-bank [Rabiner and Schafer, 1978]. On the other hand, when a model of the speech signal is applied, the parameter estimation of this model permits to better draw the characteristics of the speech signal. An important example of the parametric modeling techniques is the linear prediction [Atal and Hanauer, 1971; Makhoul, 1973] which attempts to optimally model the human speech production process of the vocal tract. More details about most common signal analysis techniques can be found in [Picone, 1993; Junqua and Haton, 1996; Mariani, 2002].

The current most commonly used acoustic parameters in ASR and speaker recognition systems are the cepstral coefficients [Oppenheim and Schafer, 1968; Furui, 1981a]. The advantage of using cepstral coefficients is their ability to separate the excitation source from the vocal tract effectively. The vocal tract configurations are very useful speaker features for characterizing speaker identity and can be well represented by cepstral coefficients. Hence, current speaker

recognition systems widely exploit these cepstral coefficients. The source-filter separation is realized by a homomorphism which transforms the convolution of the speech signal in the time domain into an addition in the cepstral domain.

The cepstral coefficients can be extracted from the output of the filter-bank on Mel scale such as Mel Frequency Cepstral Coefficients (MFCC) [Davis and Mermelstein, 1980], or can also be derived from the linear prediction coding (LPC) model, called Linear Prediction Cepstral Coefficients (LPCC). The MFCC coefficients are a type of acoustic features effective for speech recognition as well as speaker recognition applications. LPC is one of the most used techniques for low-bit rate speech coding and a very powerful method of speech signal analysis. For MFCC and LPCC coefficients, their capabilities of characterizing a speaker's voice by representing the vocal tract feature determine that both set of coefficients are suitable for the speaker diarization task as well. In the Rich Transcription Fall 2004 (RT-04F) speaker diarization evaluation [NIST, 2004] for Broadcast News (BN) data, almost all the participating systems used either MFCC coefficients or LPCC coefficients [Barras *et al.*, 2004; Tranter *et al.*, 2004; Wooters *et al.*, 2004; Moraru *et al.*, 2004b]. The subsequent part of this section is focused into presenting MFCC and LPCC parameterizations as well as differential coefficients.

A. Mel Frequency Cepstral Coefficients (MFCC)

MFCC coefficients are the cepstral coefficients obtained from the energy of a filter-bank on Mel frequency scale. Considering that the auditory perception of the human ear isn't linearly correlated to the acoustic frequency, the Mel scale [O'Shaughnessy, 1987] was introduced to map the absolute frequency of a sound into a perceptually meaningful frequency scale. This scale is designed to be approximately linear up to 1000 Hz and logarithmic beyond 1000 Hz. The correspondence between the Hertz frequency and the Mel frequency is defined as

$$f_{mel} = 2595 \log_{10} \left(1 + \frac{f_{Hz}}{700}\right) \tag{2.4}$$

As shown in Figure 2.3, the MFCC coefficients are computed with the following process:

- The speech signal is first preemphasized, then a sliding window is applied to block the signal into a set of frames. Each individual frame is then multiplied with a window so as to reduce the signal discontinuities at the boundaries of the frame.
- The Fast Fourier Transform (FFT) is performed on each windowed frame. Then the modulus of the FFT is calculated and the power spectrum is obtained.
- The power spectrum is smoothed by integrating the spectral coefficients within a filter-bank which consists of M triangular bandpass frequency filters located on Mel scale. The distance between two triangular filters is typically set to 150 mels and the width of the triangular equals to 300 mels.

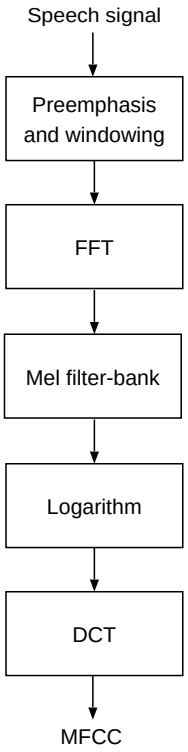


Figure 2.3: Schematic representation of MFCC parameterization.

- The log of the filter-bank output is computed so as to make the statistic of the estimated spectrum approximately Gaussian.
- In order to compress the spectral coefficients into a smaller set of coefficients and decorrelate them allowing the diagonal covariance matrices to be used in the subsequent modeling [Young, 1996], the cepstral coefficients c_n are calculated by the application of the Discrete Cosine Transform (DCT) to the log filter-bank coefficients S_k:

$$c_n = \sum_{k=1}^{M} S_k \cos\Big[n\Big(k-\frac{1}{2}\Big)\frac{\pi}{M}\Big], \qquad 0 \leqslant n \leqslant N_c \tag{2.5}$$

  where M is the number of the Mel frequency filters and N_c is the number of the cepstral coefficients to be obtained. These N_c cepstral coefficients thus constitute a cepstral vector which represents a frame of speech signal.

B. Linear Prediction Cepstral Coefficients (LPCC)

LPCC parameters are the cepstral coefficients derived from the linear prediction Coding (LPC) model [Makhoul, 1975; Markel and Gray, 1976], that is, a model simulating the speech pro-

duction process. LPC model is based on the fact that the physical characteristics of the human vocal apparatus determines the correlation between the speech samples. The LPC model exploits this correlation to reduce the amount of data without losing significant information of the vocal tract in the speech signal. The advantage of the LPC analysis is that it provides a compact and accurate representation of the speech signal with a relatively simple computation. Therefore the LPC-based parameterizations have been widely used in the ASR systems.

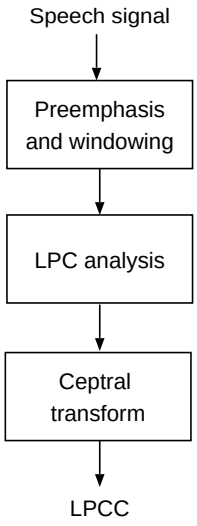


Figure 2.4: Schematic representation of LPCC parameterization.

The commonly used LPC model is the Auto Regressive (AR) model. According to the source-filter model theory, the vocal tract can be represented by an AR filter. The principle of the LPC analysis is to estimate the parameters of the AR filter (i.e. linear prediction coefficients or LPC coefficients). As shown in Figure 2.4, after blocking the preemphasized speech signal into a group of frames, a window is applied to each frame for the purpose of smoothing. Then a set of LPC coefficients is estimated for each frame of the windowed signal. There are two principal algorithms used to compute the linear prediction coefficients: the autocorrelation method [Itakura and Saito, 1968] and the covariance method [Atal and Hanauer, 1971]. Finally the LPC cepstral coefficients can be directly calculated from the LPC coefficients by the equations presented in [Rabiner and Juang, 1993]:

$$\begin{aligned} c_0 &= ln\sigma^2 \\ c_m &= a_m + \sum_{k=1}^{m-1} \Big(\frac{k}{m}\Big) c_k a_{m-k}, \quad 0 < m \leqslant p \\ c_m &= \sum_{k=1}^{m-1} \Big(\frac{k}{m}\Big) c_k a_{m-k}, \quad m > p \end{aligned} \tag{2.6}$$

where σ^2 is the gain term in the LPC model, a_m are the LPC coefficients, p is the number of the LPC coefficients calculated.

C. Differential coefficients

The cepstral coefficients discussed above are referred to as the static coefficients. In order to continuously characterize the temporal variation in the speech signal, the static cepstral coefficients are combined with some dynamic information reflecting the manner in which the coefficients vary over time. This dynamic information can be drawn by high order time derivatives (typically using first- and second-order derivatives) [Furui, 1986] of the cepstral coefficients. The delta and delta-delta parameters are the popular approximations of the first- and second-order derivatives. According to the regression analysis presented in [Furui, 1981b], the delta parameters can be calculated as:

$$\Delta c_n = \frac{\sum_{m=-k}^{k} m c_{n+m}}{\sum_{m=-k}^{k} |m|^2} \tag{2.7}$$

where the differential coefficients are obtained by implementing the regression over a window that is centered on the current coefficient c_n and covers the k preceding and k succeeding coefficients. The delta-delta parameters can be obtained recursively from the delta parameters.

2.2.2 Feature warping normalization

The static coefficients presented previously are easily corrupted by some adverse effects such as channel mismatch, additive noise and non-linear effects attributed to handset transducers. The corrupted feature parameters can't well represent speaker characteristics, thus the performance of speaker diarization systems will be degraded when using these parameters. In order to make feature parameters more robust to adverse effects, a number of feature normalization techniques have been proposed in the literature. A typical normalization method is the Cepstral Mean Subtraction (CMS) [Furui, 1981a] that is especially effective in compensating for linear channel effect. As an extension of the CMS approach, variance normalization [Koolwaaij and Boves, 2000] normalizes cepstral features by subtracting their mean and scaling by their standard deviation. Another class of normalization techniques for reducing the transmission channel effect is modulation spectrum processing such as RASTA [Hermanksy and Morgan, 1994]. However the features obtained after both CMS and RASTA methods are not robust to additive noise. Recently, a more robust normalization method was reported in [Pelecanos and Sridharan, 2001], called feature warping, which maps the distribution of an observed cepstral feature stream to a target distribution over a specified time interval. For speaker recognition, it has been found that feature warping is more robust to linear channel effects and slowly varying additive noise compared to standard normalization techniques [Pelecanos and Sridharan, 2001].

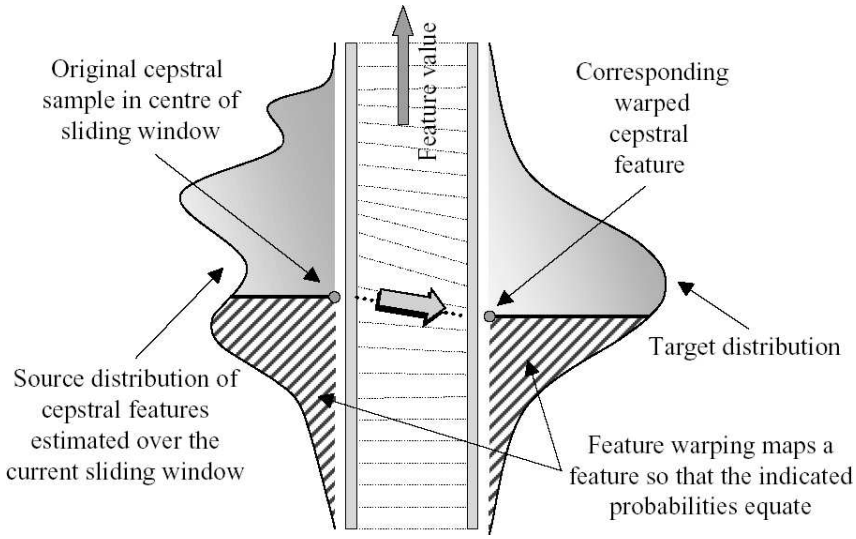


Figure 2.5: Warping of the source feature distribution to a target distribution (after Pelecanos, 2001).

As can be seen intuitively in Figure 2.5, conforming the shape of the original feature distribution to that of a standard distribution such as the normal distribution is a good method to improve the robustness of the features to linear channel effects and additive noise. This can be achieved by warping each initial cepstral feature to match the standard normal distribution. Feature warping technique is based on the assumption that the components (i.e. cepstral features) of the cepstral vector are independent. The speech signal is first parameterized into a set of cepstral vectors, then cepstral features are processed individually in their own streams of features. The warping process is performed within a sliding window covering N features. Let r be the rank of the cepstral feature centered in the middle of the current window, the corresponding warped value w can be calculated by the following equation:

$$\frac{r-\frac{1}{2}}{N} = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) dx \tag{2.8}$$

The sliding window is then shifted by a cepstral feature and the warping process is repeated until the end of the signal is reached. Feature warping technique has been shown to outperform standard normalizations for speaker verification [Xiang *et al.*, 2002; Barras and Gauvain, 2003]. The experiments reported in [Zhu *et al.*, 2005; Sinha *et al.*, 2005] have demonstrated that the feature warping approach is also effective for speaker diarization applications. This efficiency derives from the robustness of feature warping to linear channel effects and additive noise occurring in audio data. For example, in Broadcast News diarization applications, a journalist may appear at different time intervals in the same BN program to report different news, however, the reports from the same journalist may be transmitted over telephone or studio channel and the background environment of each report can vary largely. In such a case, reducing the channel

effect and additive noise will improve the performance of speaker clustering stages in diarization systems.

2.2.3 Statistical speaker modeling

The objective of speaker modeling is to characterize speaker features so as to discriminate one speaker from other speakers. The choice of good performance speaker models is a key issue in speaker diarization. The most commonly used probabilistic speaker models in speaker diarization are unimodal Gaussian [Chen and Gopalakrishnan, 1998b; Moh *et al.*, 2003] and Gaussian Mixture Model (GMM) [Meignier *et al.*, 2001; Ajmera and Wooters, 2003; Barras *et al.*, 2004]. Speaker modeling based on GMMs was applied to text-independent speaker identification by Reynolds [Reynolds, 1992] and is one of the most popular speaker models used in speaker recognition domain. Another well known speaker modeling technique for text-dependent speaker recognition applications is Hidden Markov Model (HMM) [Zheng and Yuan, 1988] that exploits the temporal characteristics of speakers. In the case of speaker diarization, an evolutive HMM approach was proposed by [Meignier *et al.*, 2001] for performing jointly the speaker segmentation and clustering.

The speaker modeling methods based on both GMMs and HMMs have been proven to be effective for speaker recognition. It exists also many other modeling techniques such as vector quantization (VQ) codebook model [Soong *et al.*, 1985], Support Vector Machine (SVM) [Schmidt, 1996] and artificial neural networks [Rudasi and Zahorian, 1991]. As Gaussian Mixture Model and Hidden Markov Model have been widely used for speaker diarization, the rest of this subsection will be focused to describe these two statistical models.

A. Gaussian Mixture Models (GMMs)

Gaussian Mixture Model is built from the given acoustic vectors by a weighted linear combination of several unimodal Gaussians. For an acoustic vector x, the mixture density is defined as

$$p(x|\Theta) = \sum_{i=1}^{M} w_i p_i(x|\theta_i) \tag{2.9}$$

where M is the number of the Gaussian components, p_i is a component density parameterized by θ_i and w_i is the corresponding component weight. The density function of each multivariate Gaussian component is given by

$$p_i(x|\theta_i) = \frac{1}{2\pi^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\Big) \tag{2.10}$$

where d is the dimension of the acoustic vector sample, μ_i is the mean vector and Σ_i is the covariance matrix. Therefore each Gaussian component is parameterized by $\theta_i = (\mu_i, \Sigma_i)$ and

the parameters of the Gaussian mixture can be denoted as

$$\Theta = (w_i, \mu_i, \Sigma_i), \quad i = 1, \cdots, M \tag{2.11}$$

with the constraint that $\sum_{i=1}^{M} w_i = 1$.

Various kinds of the GMM differ in the definition of the covariance matrix. A GMM-based speaker model can use an individual covariance matrix for each Gaussian component or an unique covariance matrix for all Gaussian components, called tied covariance. In addition, under the assumption of the independence between the acoustic features, a full covariance matrix reduces to a diagonal matrix. A typical GMM form used for speaker modeling is one where each Gaussian component has its own diagonal matrix. It has experimentally been observed that this GMM form provides better speaker recognition performances. As discussed in [Reynolds *et al.*, 2000], the good performances of using diagonal matrices are mainly derived from the fact that characterizing the distribution of the observed acoustic features by a model with a full covariance matrix can be also well obtained by using a model with a higher-order diagonal matrix, as well as the computational efficiency of diagonal matrices.

Using GMM to model speaker characteristics is motivated by two interpretations [Reynolds and Rose, 1995]:

- The Gaussian components are capable of modeling the speaker-dependent acoustic classes that represent some large phonetic classes. These acoustic classes characterize the vocal tract configurations pertaining to a specific speaker. Therefore a Gaussian mixture can be used for speaker modeling. Since the speech data used for training a speaker model is not labeled into phonetic classes, the GMM is constructed in an unsupervised manner where the classes of the observations are unknown. Hence the acoustic classes represented by a GMM are the underlying sound classes modeled from the given training data, but they need not to directly correspond to the phonetic classes.

- The second motivation for using GMM to represent speaker identity comes from the fact that a mixture model composed of enough Gaussian functions is capable of characterizing a large class of densities. When applying the GMM to speaker modeling, each Gaussian component represents the distribution of samples from one broad phonetic class, thus, the combination of these discrete Gaussians can represent the distribution of arbitrary acoustic samples from a specific speaker. This is the reason why the Gaussian mixture models are especially effective for text-independent speaker recognition applications.

A.1 Maximum likelihood parameter estimation using EM algorithm

The fundamental issue of the GMM-based speaker modeling is to estimate the parameters of the Gaussian mixture model from the given training data. A widely used parameter estimation method is the *Maximum Likelihood* (ML) estimation. Given a sequence consisting of N acoustic

vectors $X = \{x_1, \cdots, x_N\}$, and assuming that these acoustic vectors are independently drawn from the density function $p(x|\Theta)$, the GMM likelihood for the training data X is given by

$$p(X|\Theta) = \prod_{n=1}^{N} p(x_n|\Theta) \tag{2.12}$$

In the above equation, the form of the density function $p(x|\Theta)$ is known as a Gaussian mixture distribution, but the specific value of the parameter Θ (i.e. $(w_i, \mu_i, \Sigma_i),\ i = 1, \cdots, M$) is unknown. The goal of the maximum likelihood estimation is to find the parameter Θ that maximizes the likelihood of the GMM, which can be formulated as

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\ p(X|\Theta) \tag{2.13}$$

Because Equation (2.12) is a nonlinear function with respect to the parameter Θ, it is extremely difficult to optimize the parameter directly using Equation (2.13). To solve this problem, the Expectation-Maximization (EM) algorithm [Dempster *et al.*, 1977] was proposed to obtain the maximum likelihood parameter estimates. The EM algorithm is a very powerful method for estimating the parameters of a stochastic process from a given observations set. The main advantage of the EM algorithm is its capability of guaranteeing a monotonic increase in the likelihood value.

The core idea of the EM algorithm is to iteratively estimate the maximum likelihood parameters, which is the same notion as used in the Baum-Welch algorithm for the HMM parameter estimation. In the case of the parameter estimation for a GMM speaker model, the EM algorithm estimates a new GMM model parameterized by Θ_{k+1} basing on the existent GMM model parameterized by Θ_k, such that the likelihood value increases after each EM iteration, i.e. $p(X|\Theta_{k+1}) > p(X|\Theta_k)$. This iteratively parameter estimating procedure terminates when some constraints are satisfied (e.g. the likelihood increment is below a certain threshold or the maximum number of iterations is reached). In order to assure the increase of the parameter's likelihood, the Baum function has been introduced in the EM algorithm as an auxiliary function. As mathematically demonstrated in [Bilmes, 1998], the auxiliary function Q applied to the GMM parameter estimation has a form

$$Q(\Theta_{k+1}, \Theta_k) = \sum_{i=1}^{M} \sum_{n=1}^{N} \log(w_{i,k+1} p(x_n|\Theta_{i,k+1})) p(i|x_n, \Theta_k) \tag{2.14}$$

where Θ_k and Θ_{k+1} are the parameters respectively estimated at the kth and $(k+1)$th iterations, the $p(i|x_n, \Theta_k)$ is the *a posteriori* probability of the ith Gaussian component for an observed acoustic vector x_n given the existent GMM model parameterized by Θ_k. This *a posteriori* probability can be computed as

$$p(i|x_n, \Theta_k) = \frac{w_{i,k} p_i(x_n|\theta_{i,k})}{p(x_n|\Theta_k)} = \frac{w_{i,k} p_i(x_n|\theta_{i,k})}{\sum_{j=1}^{M} w_{j,k} p_j(x_n|\theta_{j,k})} \tag{2.15}$$

where $\theta_{i,k}$ is the parameters of the ith Gaussian component in the kth GMM model.

In order to obtain the parameter estimates for the new GMM model at the $(k+1)$th EM iteration, the estimations of the Gaussian component weight $w_{i,k+1}$ and the component parameter $\Theta_{i,k+1}$ are performed independently since they are uncorrelated. To find the expression of $w_{i,k+1}$, the derivation of Equation (2.14) is carried out with the constraint that $\sum_{i=1}^{M} w_i = 1$ and the introduction of the Lagrange multiplier λ. To estimate the parameter $\Theta_{i,k+1}$, Equation (2.14) is derived with respect to $\mu_{i,k+1}$ and $\Sigma_{i,k+1}$. Setting these derivatives to zero derives the following equations:

$$\begin{aligned} \sum_{n=1}^{N} \frac{1}{w_{i.k+1}} p(i|x_n, \Theta_k) + \lambda = 0 \\ \sum_{n=1}^{N} \Sigma_{i,k+1}^{-1}(x_n - \mu_{i,k+1}) p(i|x_n, \Theta_k) = 0 \\ \sum_{n=1}^{N} \Big(\Sigma_{i,k+1} - (x_n - \mu_{i,k+1})^2\Big) p(i|x_n, \Theta_k) = 0 \end{aligned} \tag{2.16}$$

where $(x_n - \mu_{i,k+1})^2$ is shorthand for $(x_n - \mu_{i,k+1})(x_n - \mu_{i,k+1})^T$.

Therefore, the new parameter estimates (i.e. $\Theta_{k+1} = (w_{i,k+1}, \mu_{i,k+1}, \Sigma_{i,k+1}),\ i = 1, \cdots, M$) for the GMM speaker model can be obtained via solving the above equation array (2.16)

$$\begin{aligned} w_{i,k+1} &= \frac{1}{N} \sum_{n=1}^{N} p(i|x_n, \Theta_k) \\ \mu_{i,k+1} &= \frac{\sum_{n=1}^{N} p(i|x_n, \Theta_k) x_n}{\sum_{n=1}^{N} p(i|x_n, \Theta_k)} \\ \Sigma_{i,k+1} &= \frac{\sum_{n=1}^{N} p(i|x_n, \Theta_k) x_n x_n^T}{\sum_{n=1}^{N} p(i|x_n, \Theta_k)} - \mu_{i,k+1}\mu_{i,k+1}^T \end{aligned} \tag{2.17}$$

Model initialization is an important factor affecting the training performance of the GMM speaker model. The GMM likelihood function has some different local maxima corresponding to the varied initial models [McLachlan, 1988]. Even if the EM algorithm can make the likelihood of a starting model converge to a local maximum likelihood, this local maximum is not ensured to be the global maximum. The choice of the initialization for GMM training is dependent on the specific applications. For example, in speaker identifications, the experiments reported in [Reynolds and Rose, 1995] have shown that a relatively simple initial model can produce speaker identification performances similar to those obtained using an elaborate initial model.

A.2 Maximum A Posteriori (MAP) adaptation from UBM

The maximum likelihood estimation can provide robust parameter estimates of a model under the condition that the amount of the training data is sufficiently large. However, in some practical applications, it is difficult to have adequately available training data. *Maximum A Posteriori* (MAP) estimation proposed by [Gauvain and Lee, 1994] has been proven to be effective to the problem of parameter estimation from sparse training data. The main advantage of the MAP estimation is that the prior information about model parameters is used in the estimation process.

The MAP method is a widely used model adaptation technique for different speech technologies. In the case of speaker modeling, a speaker model can be obtained by MAP adaptation from a Universal Background Model (UBM being a large GMM trained to present the speaker-independent distribution of features) using the training data from the speaker. In the speaker recognition domain, top performances have also been obtained by adapting a UBM via MAP estimation in the framework of NIST speaker recognition evaluations since 1996 [Reynolds *et al.*, 2000]. As presented in [Zhu *et al.*, 2005], after a first speaker clustering step, the GMM of each remaining cluster is obtained by the MAP adaptation of the means of the UBM such that the resulting cluster models to be used in a second clustering stage can better represent speaker characteristics. The LIMSI speaker diarization system incorporating this speaker clustering approach has given a state-of-the-art performance in the latest NIST Broadcast News speaker diarization evaluation [NIST, 2004]. The details about the speaker clustering technique using the MAP adaptation from UBM will be introduced later in this thesis.

The UBM is a large speaker-independent GMM (i.e. typically with 1024 or 2048 Gaussian components) trained by standard maximum likelihood approaches from a mass of data. The data used for building a UBM must be composed of the speech coming from a large number of different speakers. This diversity of the speaker components in the training data determines that the UBM is capable of modeling the speaker-independent distribution of acoustic features. The UBM can have different modes depending on the composition of the training data. According to the composition of speakers' gender, the UBM can be gender-dependent when the used training data consists only of either male speech or female speech, and it can also be gender-independent when there is a combination of male and female speech in training data. The UBM can be bandwidth-dependent or not in terms of the composition of recording channels (wideband/telephone). In applications where the gender and bandwidth information of the training speech from a speaker is a priori known, using the gender and bandwidth matched UBM can yield better parameter estimates for the speaker's model.

The main difference between MAP and ML estimation derives from the assumption that the prior distribution of the model's parameters is known. Let the sequence $X = x_1, \cdots, x_N$ be the training sample from a speaker, assuming that the prior probability density function of the parameter to be estimated is known as $p(\Theta)$, relative to Equation (2.13), the MAP estimation is defined as follows

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\ p(X|\Theta)p(\Theta) \tag{2.18}$$

When there is no a priori information concerning the distribution of the parameter Θ (i.e. the density function of the parameter $p(\Theta)$ is constant), the MAP estimation simplifies to the ML case.

The basic idea of MAP adaptation of a speaker model is to update the well-trained parameters of the UBM by using the training speech from the speaker. An example of training a speaker's model by MAP adaptation from a UBM is illustrated in Figure 2.6. As can be seen in this figure, for the components in which much training data is observed, the corresponding parameters in the speaker model largely depend on the new estimates from the training samples. On the other hand, for the components in which little training data is observed, the corresponding parameters in the speaker model heavily depend on the old estimates from the known UBM. Because UBM represents a very large acoustic space relative to the training data, some final components are acquired by just keeping the old parameters in the UBM without adaptations. More details of a variation of MAP adaptation applied to the speaker modeling are described in [Reynolds *et al.*, 2000], a summary of this adaptation approach will be given in this thesis.

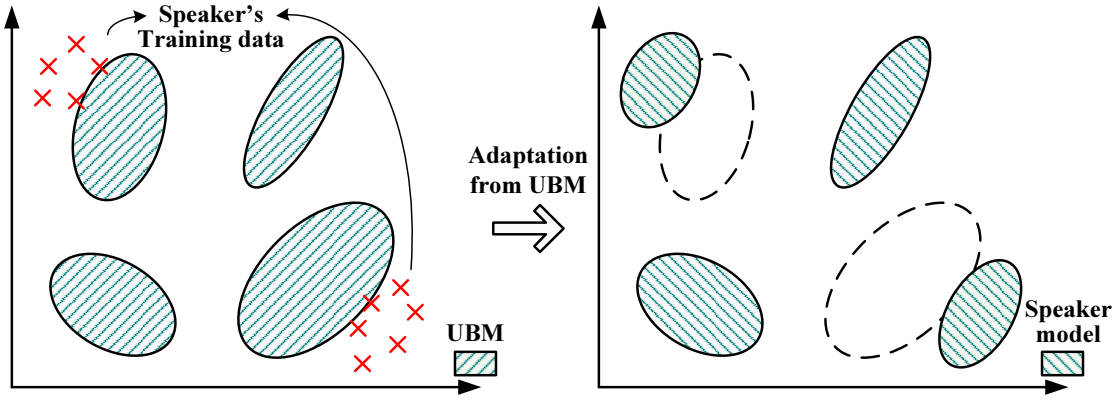


Figure 2.6: Example of the MAP adaptation from a UBM given the training data from a speaker.

In MAP adaptation of the speaker model, we first perform a probabilistic map of the training speech samples into the Gaussian components of the given UBM. For a training acoustic vector x_n, the probability of the ith Gaussian component in the UBM can be computed in terms of Equation (2.15) with replacing Θ_k by the UBM parameter Θ_{ubm}. Then, the sufficient statistics are calculated for each Gaussian component:

$$
\begin{aligned}
c_i &= \sum_{n=1}^{N} p(i|x_n, \Theta_{ubm}) \\
E_i(x) &= \frac{1}{c_i} \sum_{n=1}^{N} p(i|x_n, \Theta_{ubm}) x_n \\
E_i(x^2) &= \frac{1}{c_i} \sum_{n=1}^{N} p(i|x_n, \Theta_{ubm}) x_n x_n^T
\end{aligned}
\tag{2.19}
$$

The subsequent step of the MAP adaptation is to modify the old parameters in the UBM using the new parameters estimated from the training data according to the adaptation coefficients. Thus, the adapted means and the covariance matrices of the speaker model can be computed with the following equations:

$$\mu_{i,a} = \alpha_i E_i(x) + (1 - \alpha_i)\mu_{i,ubm} \tag{2.20}$$

$$\Sigma_{i,a} = \alpha_i E_i(x^2) + (1 - \alpha_i)(\Sigma_{i,ubm} + \mu_{i,ubm}\mu_{i,ubm}^T) - \mu_{i,a}\mu_{i,a}^T \tag{2.21}$$

where α_i is the data-dependent adaptation coefficient computed for each Gaussian component and determines the ratio between the new and the old sufficient statistics used to build the speaker model. The adaptation coefficient is given by

$$\alpha_i = \frac{c_i}{c_i + r} \tag{2.22}$$

where r is a predefined relevance factor and gives the amount of the training data that should be used to estimate the speaker model. This factor is useful to deal with the problem of balance between the new and old parameters. For Gaussian components with high probabilities of training data, $\alpha_i \rightarrow 1$ leads to the emphasis of the new parameters in the adapted speaker model. For components with low probabilities of training data, $\alpha_i \rightarrow 0$ causes the adapted model to be biased towards the old parameters. Hence, using the relevance factor makes resulting speaker model neither extremely close to the UBM nor excessively dependent on the training data.

The adaptation of the UBM presented above is carried out on both the means and the covariance matrices individually. However, in practice, it has experimentally been found that the adaptation performed only on the means of the UBM provides the best speaker verification performance [Reynolds *et al.*, 2000]. By replacing α_i in Equation 2.20 with Equation 2.22, the mean parameters can be estimated as:

$$\mu_{i,a} = \frac{c_i E_i(x) + r\mu_{i,ubm}}{c_i + r} = \frac{\sum_{n=1}^{N} p(i|x_n, \Theta_{ubm}) x_n + r\mu_{i,ubm}}{\sum_{n=1}^{N} p(i|x_n, \Theta_{ubm}) + r} \tag{2.23}$$

B. Hidden Markov Models (HMMs)

The basic theory of Hidden Markov Models was first published by Baum in the mid 1960s [Baum and Petrie, 1966]. The applications of HMMs in speech processing domain were first implemented by Baker [Backer, 1975] at CMU and by Jelinek [Jelinek, 1976] at IBM. HMMs are very important models since they have been successfully used in many state-of-the-art ASR systems. They are also effective for text-dependent speaker recognition applications where the prior knowledge about the spoken text is available. In the case of speaker diarization, performing Viterbi decoding with a HMM structure is a common segmentation technique used for partitioning the audio stream into different acoustic characteristics.

In a Hidden Markov Model, the observation is a probabilistic function of the state instead of a deterministically observable event, i.e. the state sequence is unobserved (hidden). Hence, a HMM represents a doubly embedded stochastic process where the underlying random function producing the state sequence is unobserved but can be examined by the one that causes the observation sequence. This mechanism makes HMM suitable to model the speech by representing the speech sounds and the temporal sequencing among these sounds in a single model.

Let $O = (o_1 o_2 \dots o_T)$ be a sequence consisting of T observations, a Hidden Markov Model can be formally described by the following notations:

- A set of states $S = \{s_i\}_{1 \leqslant i \leqslant N_S}$
- A set of possible observation symbols $V = \{v_k\}_{1 \leqslant k \leqslant N_V}$, each observation o_t belongs to a observation symbol
- The state transition probability distribution $A = \{a_{ij}\}_{1 \leqslant i,j \leqslant N_S}$, where a_{ij} is the transition probability from state s_i to state s_j and is defined by

$$a_{ij} = p(q_{t+1} = s_j | q_t = s_i) \tag{2.24}$$

  where q_t represents the state at time t
- The output probability distribution $B = \{b_j(k)\}_{1 \leqslant j \leqslant N_S, 1 \leqslant k \leqslant N_V}$, where $b_j(k)$ is the output distribution in state s_j and is defined as

$$b_j(k) = p(o_t = v_k | q_t = s_j) \tag{2.25}$$

- The initial state distribution $\pi = \{\pi_i\}_{1 \leqslant i \leqslant N_S}$, where $\pi_i = p(q_1 = s_i)$

In the above HMM definition, the output distribution in each state is represented by a discrete probability density when the observations are predefined as discrete symbols. However for the applications where the observations are continuous (e.g. the acoustic vectors extracted from speech signals), continuous densities are typically used as the output probabilities. For convenience, a HMM can be represented by a compact notation as $\lambda = (A, B, \pi)$. Figure 2.7 illustrates

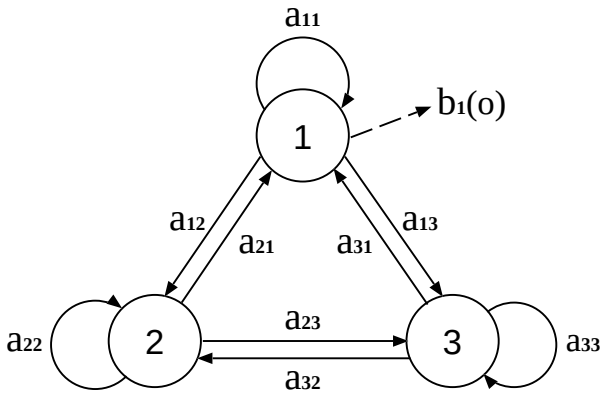


Figure 2.7: An example of 3-state ergodic HMM.

an example of 3-state ergodic HMM. In some HMMs, two special states, that are, the entry and exit states are incorporated at the start and end of models. Using these two states allows to easily connect different HMMs.

A primary application of the HMM to speaker diarization tasks is its use with the Viterbi algorithm for segmenting the audio stream in terms of different acoustic classes such as speech, silence, music and noise etc. This acoustic segmentation is usually employed to detect speech activities, which play a role as a pre-processing stage in speaker diarization systems (cf. Section 2.1). The Viterbi algorithm introduced by [Viterbi, 1967] is an effective and efficient technique for finding the optimal state sequence $q = (q_1 q_2 \dots q_T)$ in the sense that this state sequence has most likely generated the given observation sequence $O = (o_1 o_2 \dots o_T)$. Although there may be more than one most possible state sequence, the Viterbi algorithm is capable of finding the single best state path with the maximum likelihood $p(q|O, \lambda)$ (being equal to $p(q, O|\lambda)$). A crucial notion of the Viterbi algorithm is to compute the highest probability along a single path $(q_1 q_2 \dots q_t)$ where the first t observations end in state s_i at time t

$$\delta_t(i) = \max_{q_1 q_2 \dots q_{t-1}} p(q_1 q_2 \dots q_{t-1} q_t = s_i, o_1 o_2 \dots o_t|\lambda) \tag{2.26}$$

Thus we can induce $\delta_{t+1}(j)$ for all states s_j associated with the observation o_{t+1} as follows:

$$\delta_{t+1}(j) = \max_{1 \leqslant i \leqslant N_S} [\delta_t(i) a_{ij}] \cdot b_j(o_{t+1}) \tag{2.27}$$

This procedure will be performed recursively until the last observation o_T is attained. Therefore the ending state q_T can be determined by choosing the state s_i for which $\delta_T(i)$ is maximal. In order to find the single best state sequence, we track back from this ending state along the path consisting of the states which maximized Equation 2.27. The formal statement of the Viterbi algorithm can be found in [Huang *et al.*, 1990; Rabiner and Juang, 1993].

When the Viterbi algorithm is applied to the problem of acoustic classification, various structures of the HMM can be defined. For instance, the states of the HMM are typically the acoustic classes to be detected, the output probability of each state can be a GMM constructed for each desired acoustic class and the transition probabilities may be identical for all states. The Viterbi technique performed with such a HMM is the same as directly performed with a set of GMMs corresponding to the different acoustic classes. More complicated HMM structures can also be used. For example, the transition probabilities can be varied depending on the specific class transitions.

2.3 Speaker diarization

As presented in Section 2.1, the general architecture of speaker diarization systems is composed of four primary modules: front-end parameterization, speech activity detection, speaker change detection and speaker clustering. The speech parameterization has been introduced previously, three remaining modules and the common approaches employed in each module will be described in the following subsections.

2.3.1 Speech Activity Detection (SAD)

Acoustic classification is defined as the task of classifying each frame of a continuous audio stream into different acoustic classes such as speech, music, silence or background noise, etc. This process generates a set of output segments, where each segment is explicitly labeled with its corresponding acoustic class. Speech activity detection can be thought of as a basic case of the acoustic classification, which aims to classify the audio data into two broad acoustic classes: speech versus non-speech. Here, the non-speech class can be considered as a general class consisting of music, silence or background noise etc. SAD has also been called endpoint detection or Voice Activity Detection (VAD) in automatic speech detection domain.

SAD is a very useful preprocessing technique for various speech processing technologies such as automatic speech recognition, speaker identification, speaker localization and speaker diarization etc. As SAD can separate speech segments from non-speech segments, ASR systems only need to process audio segments containing speech, thus reducing the computational load. The recognition rate is highly dependent on endpoint detection, which is especially true for telephone speech where there are diverse noise derived from the speaker and the transmission system [Lamel *et al.*, 1981]. In the case of speaker identification or verification, SAD can significantly improve system performance by removing non-speech portions from training and testing signal. Speaker diarization can also benefit from SAD in the sense that the acoustic models serving the subsequent speaker segmentation and clustering modules are guaranteed to represent only speech information and not be corrupted by any non-speech data.

In the case of speaker diarization, the commonly used SAD approach is Maximum Likelihood (ML) selection with a set of models for different acoustic classes. Obviously this ML-based

method is also suitable for the acoustic classification. The acoustic classes are typically modeled by GMMs which are constructed on labeled training data, while other statistical models (e.g. HMMs or unimodal Gaussian) can also be used as the class models. The two basic acoustic classes used in ML classifier are speech and non-speech classes. The non-speech class can be split into detailed subclasses and its definition depends on the nature of the audio data to be processed. For Broadcast News data, non-speech classes mainly include music, silence or background noise classes. While for meeting data, non-speech classes are usually categorized into speech, room noise or silence classes.

For speaker diarization on BN data, various class models employed in SAD detectors can be found in the literature. Based on the discussion of [Tranter and Reynolds, 2006], the typically used class models can be classified into the following cases:

- The simplest case is to only use two class models: speech and non-speech. For example, the HMM based SAD detector employed in ICSI speaker diarization system [Wooters *et al.*, 2004] has one HMM model for speech and one HMM model for silence and music, where each HMM is composed of 3 states for enforcing the minimum duration of 30 msec and uses the same Gaussian mixtures but with the different mixture weights for each state. In addition, the SAD detector presented in [Meignier *et al.*, 2006] uses also speech and non-speech models with a set of morphological rules.

- In order to minimize the false rejects of speech in noise or music conditions (i.e. loss of speech data), additional speech models are employed in [Gauvain *et al.*, 1998; Reynolds and Torres-Carrasquillo, 2004; Barras *et al.*, 2006] which have five GMMs respectively for pure speech, noisy speech, speech over music, pure music and silence or noise. The segments labeled as noisy speech and speech over music are reclassified as speech. The use of the explicit music and noise models can also help to avoid mistakenly detecting the segments containing speech in the presence of music or noise. However the training of music and noise models is usually problematic. For music model training, the generality of the model is affected by the fact that a large number of jingles exist in BN shows and each show has its particular jingle. It is obvious that the music detection will be degraded when the test and training audio recordings come from different BN shows. For noise model training, it is very difficult to train a model adequately representing various background noise, since it is impossible to have sufficient training data which consists of all kinds of noise.

- Pure speech class can be further divided into two subclasses: wideband and narrow-band speech. For instance, the speaker diarization system developed at Cambridge university [Hain *et al.*, 1998; Sinha *et al.*, 2005] uses a GMM based SAD detector with the acoustic models for wideband speech, narrow-band speech, speech over music and pure music.

When the speech activity detection is performed on an un-segmented audio stream, there are mainly two different implementations of model-based ML classification:

- A direct implementation is to compute the likelihood of each acoustic class at every frame. When performing the speech activity detection, each frame of the input audio stream is labeled by choosing the acoustic class which has the maximum likelihood for the given acoustic vector. In MIT speaker diarization system [Reynolds and Torres-Carrasquillo, 2004], the SAD detector averages the frame likelihood over a temporal window of 50 frames and uses some heuristic smoothing rules to produce speech segments.

- An alternative implementation is a Viterbi decoder using different class models with the constraints of segment durations. In the presented Viterbi algorithm, the state at time t is only related to the state at time $t-1$, thus tending to produce very short segments. Within the framework of HMMs, a possible solution to this problem is to impose a minimum duration constraint on the class transitions by duplicating HMM states in each acoustic class (c.f. Figure 2.8). Within the framework of GMMs, the general solution is to use the transition penalty between classes in the Viterbi decoding for the purpose of creating longer segments [Gauvain *et al.*, 1998; Hain *et al.*, 1998]. The Viterbi decoding can be performed in a single pass or in an iterative fashion. As described in [Gauvain *et al.*, 1998], a Viterbi decoder is performed in a single pass to find the best state sequence which has most likely generated the observed acoustic vector sequence according to the ML criterion. While the acoustic classifier presented in [Hain *et al.*, 1998; Hain and Woodland, 1998] adapts the initial class models by Maximum Likelihood Linear Regression (MLLR) using the primary segments resulting from the Viterbi decoding, then resegments the audio stream with the adapted class models.

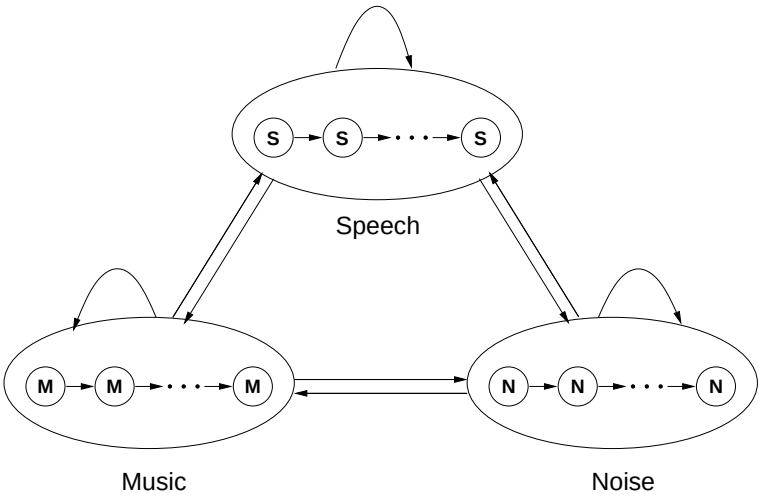


Figure 2.8: Illustration of Viterbi decoder with a 3-classes HMM topology, where the constraint of minimum segment durations can be enforced by using several intermediate states for each class.

The above SAD implementations rely on frame-by-frame approaches, since the input audio is a continuous audio stream. However, some speaker diarization systems may perform an acous-

tic change detection before the SAD stage and provide an initial segmentation to SAD. Thus, speech activity detection can be carried out on each segment individually when a preliminary segmentation is available. The acoustic change detection aims to find the points where there is an acoustic change between audio sources. These acoustic sources include particular speakers, same speaker occurring in the presence of different background noise, music, noise or signal channel characteristics etc.

For speaker diarization on meeting data, model-based ML segmentations also provide good speech activity detection performance [Fredouille and Senay, 2006; Zhu *et al.*, 2006]. However, as the non-speech comes mainly from silence or a variety of noise sources such as room noise, taps on keyboards, coughing, laughing, crosstalk, etc, this fact can be taken advantage of by energy-based SAD approaches. Anguera *et al.* have proposed to integrate an energy-based SAD algorithm into the SAD detector within the ICSI meeting speaker diarization system in [Anguera *et al.*, 2006a]. The advantage of the energy-based SAD method is that retraining the acoustic models is not needed and thus this approach is more robust to new meeting data. As introduced in [Anguera *et al.*, 2006a], a hybrid system combining an energy-based detector and a model-based decoder is used in ICSI meeting speaker diarization system. This SAD system first performs an energy-based detection of silence sections with a varied threshold automatically adapted to produce enough non-speech segments, then employs a Viterbi decoder with the acoustic models trained on the segmentation from the first stage to produce the final speech segments. This hybrid SAD system has been experimentally shown to provide better SAD performance for the meeting data in the latest NIST meeting speaker diarization evaluation [NIST, 2006].

2.3.2 Speaker change detection

In the presented diarization system architecture (c.f. Section 2.1), after the SAD stage has been executed to provide all speech sections in the audio, a subsequent speaker change detection step is then performed on each speech segment to find the time points where there is a change of speaker. This step has been also referred to as speaker segmentation. However, when the input to this step is an un-segmented audio stream, a more general acoustic change detection is required, which consists in locating segment boundaries corresponding to the changes in the acoustic characteristics. The acoustic change detection aims to locate both speaker changes and other acoustic changes among speech, silence, music and background noise classes (the identification of acoustic classes being not necessary). In addition, it also searches for the changes in a speaker turn where the background environments switch. Although acoustic and speaker change detection are applied at a different level, both of them rely on the same approaches.

At the speaker change detection step, only a primary segmentation needs to be provided and the segments are not required to be labeled in terms of speaker, since a subsequent speaker clustering which regroups the segments from the same speaker will give a speaker label to each segment, and a possible re-segmentation will refine the segment boundaries. However, pure segments are expected to be given, as they have a direct impact on the performance of the segment models which will be used in the speaker clustering step. Therefore, the main objective of speaker

change detection is to minimize missed change points so that each segment contains speech from a single speaker. At the same time, as the performance of the clustering stage depends largely on the amount of available data for each segment to be clustered, the speaker change detection needs to produce segments with sufficient long durations. The false change points produced by splitting a speaker turn into several segments are less important, since this kind of errors may be eliminated in the speaker clustering.

Speaker change detection is a difficult task due to the fact that the duration of some speaker turns may be very short. Several approaches have been proposed in the literatures. According to the classification presented in [Chen and Gopalakrishnan, 1998b; Kemp *et al.*, 2000], three classes are defined:

- Silence-based speaker change detection
- Model-based speaker change detection
- Metric-based speaker change detection

These methods can work individually or be combined with another method to detect speaker changes. Two main structures of change detection systems have been proposed in the publications. In the first system structure, change points are found by performing a detection technique in a single pass. The second structure of detection systems is a more complicated one that performs different kinds of approaches in multiple passes. Typically, the second form consists of two stages where several potential speaker changes are first proposed in a first step, then all candidating speaker changes are reestimated and some of them are validated as the final decisions of speaker changes. The multiple passes systems can also carry out one detection algorithm iteratively to optimize resulting change points. The rest of this subsection will be focused on the introduction of the different detection methods and their implementations in different systems.

A. Silence-based speaker change detection

Silence-based approaches are relatively simple methods used for speaker change detection. They assume that there is always a short period of silence between two speech segments from different speakers and a speaker turn is not interrupted by silence segments. The silence-based change detection is generally performed in two phases: silence regions in the audio are first located, then the speaker change points can be placed at the the silence regions if some additional constraints are satisfied. The silence sections can be detected either directly by an energy detector or by a decoder. In this sense, the silence-based detection approaches can be further distinguished as energy-based and decoder-based approaches.

The energy-based systems rely on the fact that the energy level of silence segments is lower than that of voiced segments. Basically, an energy detector is first employed to locate silence periods by measuring and thresholding the audio energy. As described in [Nishida and Ariki, 1998; Kemp *et al.*, 2000], the power or the mean of it is calculated every several milliseconds for the

input audio, then the regions which have their energy less than a given threshold are detected as the silence. As the threshold varies largely depending on audio type and relates closely to the speech noise ratio of the input audio, it is difficult to tune it automatically.

The decoder-based approaches use a word or phone recognizer to decode the input audio and the silence locations can be obtained from the label information produced by the decoder. This kind of detections has been implemented by [Kubala *et al.*, 1997; Woodland *et al.*, 1997; Liu and Kubala, 1999]. Some systems also make use of the additional gender informations generated by the recognizers (e.g. [Hain *et al.*, 1998]). Because the decoding is run at a phone or word level, this makes the segment generation step to over-segment the speech data, thus some additional merging is required to form speech segments with reasonable durations.

The common second step of both silence-based speaker change detections is to locate segment boundaries at the silence regions according to some extra constraints. A typical constraint is the minimum duration of silence segments, only the silence sections longer than the minimum duration are categorized as silence segments. The definition of this duration is a key issue of silence-based detection approaches. On one side, using a shorter minimum duration leads to detect the silences between phones, this word cutting will reduce recognition accuracy when the segmentation is used for speech recognition. On the other side, using a longer minimum duration tends to miss some speaker changes, this degrades the performance of speaker diarization systems. In addition, to deal with the problem of over-segmentation, some other smoothing rules are also used to relabel short segments and merge them into their neighbors [Hain *et al.*, 1998]. However, there is no definite relation between speaker changes and the existence of silence. In fact, for Broadcast News data, a segment of music or jingles often occur between the speech from different speakers; for meeting data, the speaker changes usually take place very fast, even with overlap sometimes. Therefore the silence-based detection methods are not often used to detect speaker changes independently, while it can be performed to provide a second detection stage with putative speaker changes.

B. Model-based speaker change detection

Another widely used segmentation approach relies on a set of prior models for different acoustic classes such as wideband speech, telephone speech, male speech, female speech, pure speech, pure music, silence and noise, etc. This method can be considered to an extension of the silence-based segmentations, in which a segment boundary is located at any change point between two different acoustic classes rather than one only between speech and silence. Since the model-based technique relies on the identification of the acoustic characteristics, the Viterbi decoder applied for the acoustic classification (C.F. Section 2.3.1) can be also used to perform the model-based change detection. It should be noted that the detection performance of the model-based approach depends lots of on the similarity between the training and test data. For the case of the speaker diarization on broadcast news, the BN shows used to train the models need to have the same style as of the test data. In the domain of the diarization on meetings, it is better that the training data and the test data come from the same source with a similar pattern of the organization.

Different implementations of the model-based approach can be found in the publications. The model-based segmenter presented in [Kemp *et al.*, 2000] uses a speech recognizer with four word models for anchor, field, music and silence, where each HMM state uses diagonal variance GMM with different number of mixtures. In the diarization systems of [Hain *et al.*, 1998; Tranter and Reynolds, 2004], the segmentation is carried out using a phone recognizer subsequent to the acoustic classification stage. The used models consist of 45 context independent phone model per gender, a silence/noise model and a null language model. A set of segments are thus obtained by merging the frames with the same gender label.

C. Metric-based speaker change detection

Metric-based segmentations are probably the most used method for detecting speaker changes. The basic idea of this approach is to measure the dissimilarity between two consecutive audio segments via a distance function. The distance measure can be defined by either a statistic distance between two distributions of the data within each segment or a score relying on the likelihood of the data under the model generating it. Apparently a high distance value implies that the two segments are uttered by different speakers, whereas a low value indicates that both segments come from a same speaker.

The general procedure of the metric-based segmentation approacher is illustrated in Figure 2.9, which involves two adjacent windows of the same fixed length. The distance between two segments is continually computed when the pair of windows are shifted along the audio stream. The peaks in the distance curve are determined as change points if their absolute values are above a predefined threshold. In order to reduce false change points, the distance graph is usually smoothed by a low-pass filtering and a minimal duration is imposed on two neighboring maxima.

Various metric-based segmentation algorithms differ in the choice of distance measure, windowing technique, distance smoothing and thresholding decision. The following of this section presents the different distance metrics that have been successfully used to detect speaker changes as well as their implementations.

C.1 Generalized Likelihood Ratio (GLR) metric

The GLR distance measure was first introduced into the speaker segmentation by Gish in 1991 [Gish *et al.*, 1991]. This likelihood-based metric is computed as a likelihood ratio between two following hypotheses:

- H_0: both segments $X_1 = \{x_1, \cdots, x_i\}$ and $X_2 = \{x_{i+1}, \cdots, x_N\}$ are produced by a same speaker $X = X_1 \bigcup X_2 \sim M(\mu, \Sigma)$
- H_1: the segments X_1 and X_2 are produced by different speakers $X_1 \sim M(\mu_1, \Sigma_1)$ and $X_2 \sim M(\mu_2, \Sigma_2)$

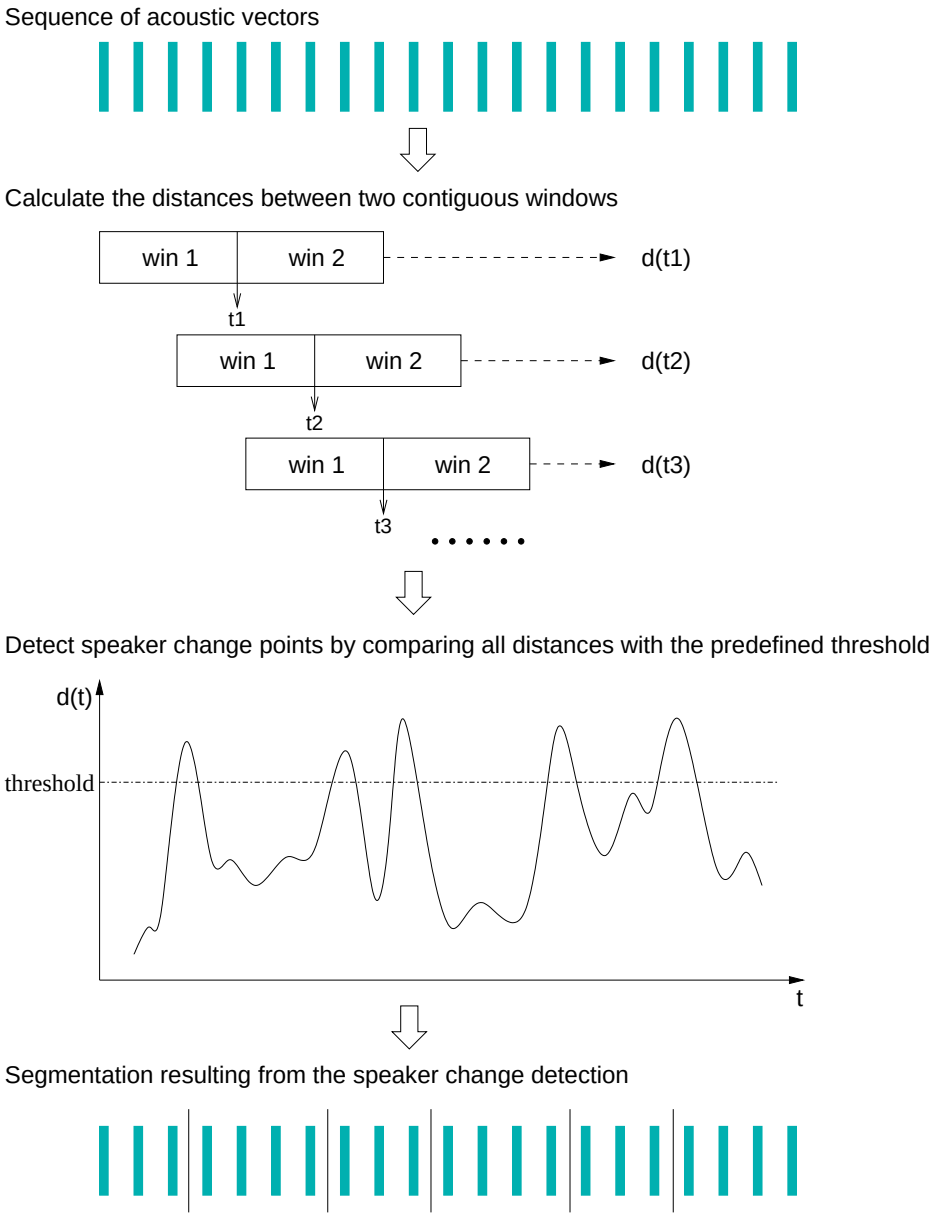


Figure 2.9: Illustration of a metric-based speaker change detector using two adjacent sliding windows.

Thus, the corresponding likelihood ratio is defined as:

$$R = \frac{L(X, M(\mu, \Sigma))}{L(X_1, M(\mu_1, \Sigma_1))L(X_2, M(\mu_2, \Sigma_2))} \tag{2.28}$$

where the $L(X_i, M(\mu_i, \Sigma_i))$ is the likelihood of the segment X_i given the model $M(\mu_i, \Sigma_i)$.

Consequently, $L(X, M(\mu, \Sigma))$ stands for the likelihood of the two segments being uttered by the same speaker and $L(X_1, M(\mu_1, \Sigma_1))L(X_2, M(\mu_2, \Sigma_2))$ represents the likelihood of each segment being uttered by the different speaker. To obtain a distance between two segments, the logarithm of this likelihood ratio is taken:

$$GLR(X_1, X_2) = -\log R \tag{2.29}$$

Then this GLR distance is compared with a threshold to decide whether both segments are spoken by the same speaker or not. The GLR is a very similar metric to the conventional Log Likelihood Ratio (LLR). The main difference between both ratios exists in the estimation of the unknown model parameters: the models involved in GLR are constructed directly on the data within each considered segment, while the model parameters for LLR are estimated a priori from the training data.

In the case of that each segment is modeled by a multivariate Gaussian process, the GLR metric can be expressed as (the related mathematical demonstration can be found in Appendix A):

$$GLR(X_1, X_2) = \frac{N}{2}\log|\Sigma| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| \tag{2.30}$$

where Σ is the covariance matrix for the union of two segments, Σ_1 and Σ_2 for each segment X_1 and X_2 respectively, and N_1 and N_2 are respectively the count of the data samples in the segment X_1 and X_2.

The GLR has been widely used as the distance measure for detecting speaker changes in different applications. In the field of speaker diarization for broadcast news, a GLR speaker change detector is integrated into the step-by-step diarization system of [Meignier *et al.*, 2006]. For broadcast news transcription and indexing, Liu and Kubala propose a two-stage detection algorithm that runs a gender-independent phone-class decoder (i.e. classifying typical phone models used for ASR into several broad phone groups) as the first pass and verifies the speaker change point between every output phone with using a modified GLR metric based speaker segmentation. In order to diminish the bias that more data within the segment results in a higher GLR value, a penalty factor is introduced into the likelihood ratio as described in [Liu and Kubala, 1999]:

$$R' = \frac{R}{(N_1 + N_2)^\theta} \tag{2.31}$$

where θ is a empirically determined parameter. This penalized GLR metric is similar to the Bayesian Information Criterion (BIC) in which the penalty term takes into account the model complexity besides the size of acoustic data (the details about the BIC metric can be found in the following subsection). It is experimentally found that the penalized GLR distance is more robust than the standard GLR [Liu and Kubala, 1999].

In two-speaker conversational speech or speaker recognition fields, some speaker change detection methods based on the GLR distance have been presented in literatures. The speaker change

detector explained in [Adami *et al.*, 2002] assumes the beginning second speech from the conversation uttered by the first speaker and selects the speech segment with the maximum of the GLR distance from the first second data to represent the other speaker. The second step is to locate the speaker change boundaries at the time points where the GLR distances from both speakers are equal and each segment is assigned to the speaker having a lower distance value.

In [Gangadharaiah *et al.*, 2004], another speaker detection algorithm is also proposed for the two-speaker segmentation task. The speaker change candidates are first found out by the GLR-based speaker segmentation using an automatic threshold setting technique that was originally introduced in [Lu and Zhang, 2002] (see Subsection C.3 for more details). After training the two speaker models on the selected data from the speech segments produced by the first stage, a Viterbi decoding with equal state transition probabilities is performed to label each segment with the more likely speaker and refines the speaker change locations.

To avoid the use of the tunable threshold that is largely dependent on the type of data, a robust threshold technique is investigated in the two-pass segmenter (c.f. called DISTBIC) presented in [Delacourt and Wellekens, 1999; Delacourt *et al.*, 1999a; Delacourt *et al.*, 1999b; Delacourt and Wellekens, 2000], where the GLR-based potential speaker changes detection is followed by the BIC-based refinement algorithm. In place of directly using the absolute GLR values, a low-pass filtering is first performed to smooth the distance curve, then a change point is detected by a local maximum GLR value if the differences between this maximum and its surrounding minima are above a certain threshold defined as a proportion of the distance variance. In the work of [Delacourt and Wellekens, 1999; Delacourt and Wellekens, 2000], some other distance measures (e.g. the Kullback Leibler metric) are also used in the first step of this two-stage segmentation. In spite of the high computational cost of the GLR distance, it is proven to be more effective for the speaker change detection than the other metrics since it can provide sharp peaks at true speaker change points.

C.2 Bayesian Information Criterion (BIC) metric

The Bayesian Information Criterion (BIC) is probably the most dominant dissimilarity measurement used in metric-based segmentation approaches. The BIC criterion introduced in [Schwarz, 1978] is a model selection criterion which maximizes the likelihood of the model penalized by the model complexity. It is inspired by the penalized likelihood principle of AIC (Akaike's Information Criterion) [Akaike, 1974] and is also known as the Minimum Description Length (MDL). In detail, given a sequence consisting of K observation samples $X = \{x_1, \cdots, x_K\}$ and a set of models $\mathcal{M} = \{M_i : i = 1, \cdots, m\}$, the BIC criterion of each model M_i is defined as :

$$BIC(M_i) = \log L(X, M_i) - \lambda \frac{1}{2} \#(M_i) \log K \tag{2.32}$$

where $L(X, M_i)$ is the maximum likelihood of the data set X under the model M_i, $\#(M_i)$ is the number of the free parameters in the model M_i, λ is the penalty weight and set to 1.0 in theory, but in practice, it needs to be tuned to the data. The BIC model selection method uses the above

penalized likelihood function to reflect how well the model M_i fits the given data and choose the model with the maximum BIC value to describe it.

When the BIC criterion is used to determine speaker changes, the question of whether a change occurs between two neighboring audio segments can be considered as a model selection problem between two following models:

- a single model M generates both segments $X_1 = \{x_1, \cdots, x_i\}$ and $X_2 = \{x_{i+1}, \cdots, x_N\}$
- two different models M_1 and M_2 generate the segments X_1 and X_2 respectively

Thus, the difference between the BIC values of these two models can be written as:

$$\Delta BIC(i) = \log R - \lambda \frac{1}{2} \Delta \#(M, M_1 + M_2) \log N \tag{2.33}$$

where R is the likelihood ratio of the test that two segments are uttered by the same speaker or different speakers (c.f. Equation 2.28), $\Delta \#(M, M_1 + M_2)$ is the difference between the number of free parameters in the single model and the two individual models and λ is the penalty weight. A negative value of ΔBIC indicates that the data is better drawn from two distinct models than a single model, thus a speaker change point can be located at the boundary between two audio segments.

Under the assumption of each model composed of a single Gaussian, the Equation 2.33 is expressed as:

$$\Delta BIC(i) = -\frac{N}{2} \log |\Sigma| + \frac{N_1}{2} \log |\Sigma_1| + \frac{N_2}{2} \log |\Sigma_2| + \lambda P \tag{2.34}$$

For the case of a full covariance matrix, the penalty function P is given by

$$P = \frac{1}{2}(d + \frac{d(d+1)}{2}) \log N \tag{2.35}$$

where d is the dimension of the acoustic feature space.

The BIC criterion was first applied for the task of speaker change detection by Chen and Gopalakrishnan [Chen and Gopalakrishnan, 1998b]. The proposed multiple changes detection method involves a sliding variable-size analysis window and evaluates each frame within the current window via calculating the corresponding ΔBIC value between the left and right subsets of audio around the examined frame. When several frames have negative values of ΔBIC in a same window, the change point is located at the frame with the minimum ΔBIC. If a change point is detected, the analysis window is reset to the change point and the search process repeated. If no change is detected, the current window is increased by a predefined amount of data and the ΔBIC is computed for each frame within the new window. This BIC change detection algorithm has been broadly adopted in different speaker segmentations with some improvements.

The original BIC algorithm of [Chen and Gopalakrishnan, 1998b] has been shown to provide good detection performance but it does not perform well on short segments that have the duration less than 2 seconds. This is because the accuracy of the estimated Gaussian models depends largely on the amount of data available in each audio segments. To address the issue of inadequate data, a new window selection technique is proposed in [Tritschler and Gopinath, 1999], where the search window is enlarged by a large amount of data when a change point is detected recently. In addition the MAP-adapted BIC segmentation of [Roch and Cheng, 2004] also demonstrated its capability of detecting changes between small segments, in which the models are obtained by MAP adaptation from the prior model trained on a pool of individual speakers. Recently, a more robust BIC based segmentation is proposed in [El Khoury *et al.*, 2007]. The first step is to detect a putative change point within each uniform window of 2 seconds according to the GLR or BIC metric. Based on each found change point, a new window is built between its precedent and subsequent detected points and the search is repeated within each new window. Two close output points with interval less than 0.2 seconds are replaced by their mean. This procedure is repeated until no more point change is possible. Finally, the BIC criterion is used to confirm all change candidates.

Another disadvantage of the BIC detection algorithm is its high computational cost. Some ways have been proposed to make the BIC computation more efficient. In [Tritschler and Gopinath, 1999], the ΔBIC is not computed for the frames close to the beginning of a large search window. The speaker change detector described in [Cettolo, 2000; Cettolo *et al.*, 2005] searches change candidates via computing ΔBIC values with a low resolution rate and validates the potential changes by computing ΔBIC values on the window centered around the candidate using a higher resolution rate. The most used approaches to reduce the BIC computation is the use of a hybrid structure in segmentation systems where the BIC metric is used in a second-stage to confirm the putative changes produced by a faster change detector. Various distance metrics have been used in a first speaker change step to provide potential changes: the GLR distance in the DISTBIC segmentation of [Delacourt and Wellekens, 1999; Delacourt *et al.*, 1999a; Delacourt *et al.*, 1999b; Delacourt and Wellekens, 2000], a normalized log-likelihood ratio in [Vandecatseye and Martens, 2003], the Hotelling's T^2 measure in [Zhou and Hansen, 2000; Reynolds and Torres-Carrasquillo, 2004; Tranter and Reynolds, 2004] and the Divergence Shape Distance (DSD) in [Lu *et al.*, 2001; Lu and Zhang, 2002].

Although BIC-based segmentations avoid the use of an explicit threshold that exists in most metric-based segmentations such as the GLR, the penalty weight λ introduced in the BIC formula is indeed an implicit threshold and has to be adjusted according to the data. Therefore the tuning of this empirical factor becomes a critical issue of the BIC change detector. It has been reported in [Delacourt and Wellekens, 2000] that the optimal value of λ is dependent on the type of the data. A modified BIC criterion without the need of the factor λ has been presented in [Ajmera *et al.*, 2002], where each of the two individual models M_1 and M_2 is composed of a single Gaussian and the combined model M consists of two Gaussian components. The equal complexities between the two distinct models and the single model result in the cancellation of the penalty term in the BIC formulation.

C.3 Kullback Leibler (KL) distance

The Kullback Leibler (KL) distance is also a successfully used metric for the speaker change detection [Siegler *et al.*, 1997]. The KL divergence (also being referred to as Relative Cross Entropy) is a measure of the differences between two probability distributions:

$$KL(A,B) = E_A\Big(\log P(A) - \log P(B)\Big) \tag{2.36}$$

where E_A is the expectation with respect to the Probability Distribution Function (PDF) of the observation A (denoted as $P(A)$). In the case of the speaker change detection, two audio segments X_1 and X_2 can be considered as the two distributions A and B respectively. Assuming that the acoustic vectors within each segment conform to a Gaussian distribution, the KL distance between two segments can be formulated as (c.f. [Campbell, 1997]):

$$KL(X_1,X_2) = \frac{1}{2}tr[(\Sigma_1-\Sigma_2)(\Sigma_2^{-1}-\Sigma_1^{-1})] + \frac{1}{2}tr[(\Sigma_2^{-1}-\Sigma_1^{-1})(\mu_1-\mu_2)(\mu_1-\mu_2)^T] \tag{2.37}$$

where tr is the trace function of square matrix.

To obtain a symmetric distance metric, the KL2 distance is defined as [Siegler *et al.*, 1997]:

$$KL2(X_1,X_2) = KL(X_1,X_2) + KL(X_2,X_1) \tag{2.38}$$

Under the same Gaussian assumption mentioned above, the KL2 can be computed by the following equation [Delacourt and Wellekens, 2000]:

$$\begin{aligned} KL2(X_1,X_2) &= \frac{1}{2}(\mu_2-\mu_1)^T(\Sigma_1^{-1}+\Sigma_2^{-1})(\mu_2-\mu_1) \\ &+\frac{1}{2}tr[(\Sigma_1^{1/2}\Sigma_2^{-1/2})(\Sigma_1^{1/2}\Sigma_2^{-1/2})^T] \\ &+\frac{1}{2}tr[(\Sigma_1^{-1/2}\Sigma_2^{1/2})(\Sigma_1^{-1/2}\Sigma_2^{1/2})^T] - d \end{aligned} \tag{2.39}$$

where d is the dimension of acoustic vectors. The speaker segmentation method described in [Siegler *et al.*, 1997] is based on the KL2 distances between two adjacent windows shifted along the audio stream. In [Delacourt *et al.*, 1999b; Delacourt and Wellekens, 2000], the KL2 distance is used in the first step of the two-passes speaker change detector.

As the sample mean is easily affected by environment conditions or transition channels, the second term of the KL distance in Equation 2.37 can be ignored in practice, thus bringing out the Divergence Shape Distance (DSD) as a variation of the KL distance:

$$DSD(X_1, X_2) = \frac{1}{2} tr[(\Sigma_{X_1} - \Sigma_{X_2})(\Sigma_{X_2}^{-1} - \Sigma_{X_1}^{-1})] \tag{2.40}$$

The two-stages unsupervised speaker change detection system presented in [Lu *et al.*, 2001; Lu and Zhang, 2002] implements a DSD-based speaker segmentation as the first step, in which an adaptive threshold is obtained from the previous K consecutive DSD distance:

$$Th_i = \alpha \frac{1}{K} \sum_{k=0}^{K} DSD(i-k-1, i-k) \tag{2.41}$$

where α is an amplification coefficient. Thus a potential speaker change is detected between two segments when their DSD distance value is greater than two neighboring distances around it (i.e. the preceding and succeeding ones) and the automatically computed threshold as well. The segment boundaries found in the first step are verified and refined via the BIC metric.

Some improvements to the above two-stages speaker segmentation have been reported in [Wu *et al.*, 2003a; Wu *et al.*, 2003b]. In the first segmentation step, acoustic vectors are first classified into three categories relying on the likelihood of each frame given a pre-trained speaker-independent UBM, then only the speech frames with high likelihood probabilities are used to calculate the DSD metric between two adjacent segments. Moreover, instead of the BIC-based refinement approach, the potential speaker changes are confirmed according to the likelihood of the current segment to the last speaker model trained on the previous segments.

C.4 Other distance measures

Apart from the above described dissimilarity metrics, some other distance measures are also successfully used to perform speaker change detection. The two-stage segmentations of [Zhou and Hansen, 2000; Reynolds and Torres-Carrasquillo, 2004; Tranter and Reynolds, 2004] commonly adopt the Hotelling's T^2 distance in the first detection stage:

$$T^2(X_1, X_2) = \frac{N_1 N_2}{N} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \tag{2.42}$$

In [Hung *et al.*, 2000], three different dissimilarity measures (i.e. the KL, Mahalanobis and Bhattacharyya distances) are implemented in the same segmenter and are experimentally shown to give comparable change detection performances. The Mahalanobis and Bhattacharyya distances are respectively defined as:

$$M(X_1, X_2) = \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma_1^{-1} \Sigma_2^{-1} (\mu_1 - \mu_2) \tag{2.43}$$

$$B(X_1, X_2) = \frac{1}{4} (\mu_1 - \mu_2)^T (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \log \frac{|\Sigma_1 + \Sigma_2|}{2|\Sigma_1 \Sigma_2|^{\frac{1}{2}}} \tag{2.44}$$

In the work of [Gauvain *et al.*, 1998; Barras *et al.*, 2004], a geometric weighted version of Mahalanobis distance (referred as Gaussian divergence measure) is used to detect acoustic changes in the audio and can be formulated as

$$G(X_1, X_2) = (\mu_1 - \mu_2)^T \Sigma_1^{-1/2} \Sigma_2^{-1/2} (\mu_1 - \mu_2) \tag{2.45}$$

The metric-based segmentation presented in [Kemp *et al.*, 2000] uses the entropy loss to code the two contiguous segments separately and compares it with the GLR and KL distances. In [Delacourt and Wellekens, 1999; Delacourt and Wellekens, 2000] some other second-order statistical measures are also investigated in comparison of the GLR and KL distances and the GLR distance is experimentally proven to be most effective in the proposed two-stage segmentation system.

More recently, a cross probability measure, called as XBIC is designed by the comparison of the BIC criterion and a distance for HMMs [Anguera, 2005]. The XBIC segmentation technique uses two adjacent windows shifted through the audio and detect acoustic changes via calculating the cross probabilities of each segment against the model trained on the data within the other segment:

$$XBIC(X_1, X_2) = L(X_1, M_2) + L(X_2, M_1) \tag{2.46}$$

Similar to the XBIC measure of [Anguera, 2005], a normalized bilateral score is proposed by the authors of [Malegaonkar *et al.*, 2006], which introduces the probabilities of the data from each segment for compensating the speech variations from the same speaker.

The speaker change detection of [Mori and Nakagawa, 2001] exploits a Vector Quantization (VQ) method that creates a codebook using the data from one of the two adjacent segments by a VQ algorithm and measures the dissimilarity via the VQ distortion (c.f. [Nakagawa and Suzuki, 1993]) between the codebook and the other segment. This measure has been experimentally proven to be more robust for short segments than the GLR and BIC metrics.

2.3.3 Speaker clustering

After performing the speaker segmentation to generate a set of distinct homogeneous segments (i.e. each segment being assume to present one single speaker in a particular background and channel condition), the following step of a diarization system is to to group together the segments that come from the same speaker (c.f. Figure 2.1). An ideal speaker clustering is capable of generating one cluster for each speaker occurring in the audio, with all segments from a given speaker associated with a single cluster. In fact, the speaker clustering is a general task and the clustering output may be useful for some other speech processing applications than the speaker diarization. For example, clustering results are beneficial for improving ASR performance in the sense that the pre-trained acoustic models can be adapted to a new speaker using the speech data within the cluster corresponding to the new speaker. For the speaker recognition application

where a set of unlabeled utterances need to be partitioned into clusters in terms of speaker attributes, the speaker clustering is performed independently without a precedent change detection stage.

In the context of this thesis, the speaker clustering task is assumed to be a unsupervised classification problem, where both the number of speakers and the speaker identities are unknown a priori. Additionally, since the speaker diarization task is usually relevant to a given audio, the clustering stage produces recording-internal speaker labels rather than true speaker identities. However, it is possible to have prior knowledge in some specific applications. The prior information could be the number of speakers involved in an audio recording such as telephone conversation or regular group meeting, or training data available for some speakers such as the anchors in particular news shows or meeting participants in successive group meeting recordings. The effect of using different a priori information for speaker diarization is investigated in the work of [Moraru *et al.*, 2004a; Moraru, 2004]. It is shown that using sufficient amount of speaker training data can improve diarization performance but knowing the number of speakers seems to be not very helpful.

Many speaker clustering methods have been proposed in the bibliography, which can be classified into two categories according to their application fields: on-line and off-line speaker clustering. The main difference between them derives from the different levels of access to the data that needs to be processed. The clustering algorithms applied to on-line applications are required to made clustering decisions immediately once a new audio segment or a batch of data is received. This means the complete data information is no available for the on-line speaker clustering. On the contrary, off-line clustering techniques are allowed to access all the data before processing it.

The following part of this section reviews the speaker clustering methods used for either on-line or off-line applications, but it is focused on the introduction of the off-line clustering approaches due to their extensive utilization. The clustering techniques to be presented here aim to deal with mono-channel audio recordings. There are some other clustering approaches relying on the use of the multi-channel information in meeting recordings [Pardo *et al.*, 2006a; Pardo *et al.*, 2006b; Gallardo-Antolin *et al.*, 2006], but they are not concerned in this thesis.

A. Sequential clustering techniques

Sequential clustering approaches are widely used for the on-line applications where no information on all the data is available. The basic idea of the sequential clustering technique is illustrated in Figure 2.10. At the beginning of the clustering, the first received segment is considered as a cluster. When each new segment arrives, it is assigned to either an existent cluster or a new cluster according to a merging criterion. Since this clustering algorithm processes input segments in temporal order, a stopping criterion is not necessary for determining the optimum number of clusters, thus reducing the computational complexity of the clustering stage.

The sequential speaker clustering presented in [Mori and Nakagawa, 2001] uses a Vector Quantization (VQ) distortion algorithm as the distance metric. A new segment is merged into one

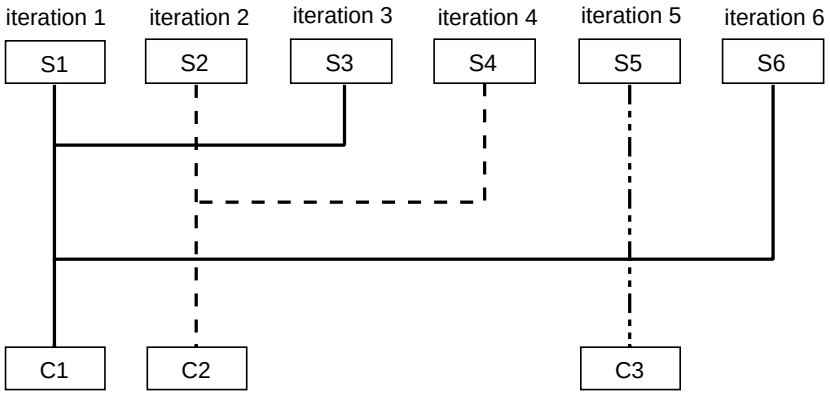


Figure 2.10: Example of the sequential clustering approaches.

present cluster that has the minimum VQ distortion with this segment, if their VQ distortion is below a given threshold; otherwise a new cluster is created as a new codebook.

In [Liu and Kubala, 2003], several on-line clustering algorithms are introduced and compared, all of which fall into the sequential speaker clustering framework. The leader-follower clustering exploits a K-means algorithm to measure the similarity between a new segment and all existent clusters and uses a predefined threshold to decide whether the new segment is combined into the nearest cluster or not. In order to avoid the use of data-dependent thresholds, the authors also propose to use the GLR as the distance measure for selecting the nearest cluster and make the merging decision via the within-cluster dispersion that was first proposed in [Jin *et al.*, 1997] (more details will be given in the next subsection). Since it is found that using dispersion alone is not very effective for speaker clustering, a hybrid system that combines GLR upper and lower thresholds with the within-cluster dispersion is thus proposed in the same work of [Liu and Kubala, 2003]. The leader-follower clustering and the hybrid clustering system have shown to outperform the dispersion-based clustering technique.

B. Hierarchical clustering techniques

In the case of off-line processing, most speaker clustering systems presented in the literature rely on a conventional hierarchical scheme where the optimum clustering is obtained from a dendrogram consisting of possible clustering solutions. Figure 2.11 illustrates an example of the hierarchical clustering. In this kind of approaches, a tree of clusters is constructed by iteratively merging or splitting clusters and the best clustering solution is chosen at the place where the optimum number of clusters is reached.

There are two main approaches to perform hierarchical clustering: agglomerative (also called bottom-up) method and divisive (also called also top-down) method. The agglomerative clustering starts with a set of speech segments where each segment is considered as an initial cluster and recursively merges nearest clusters until a stopping criterion is attained. The divisive clus-

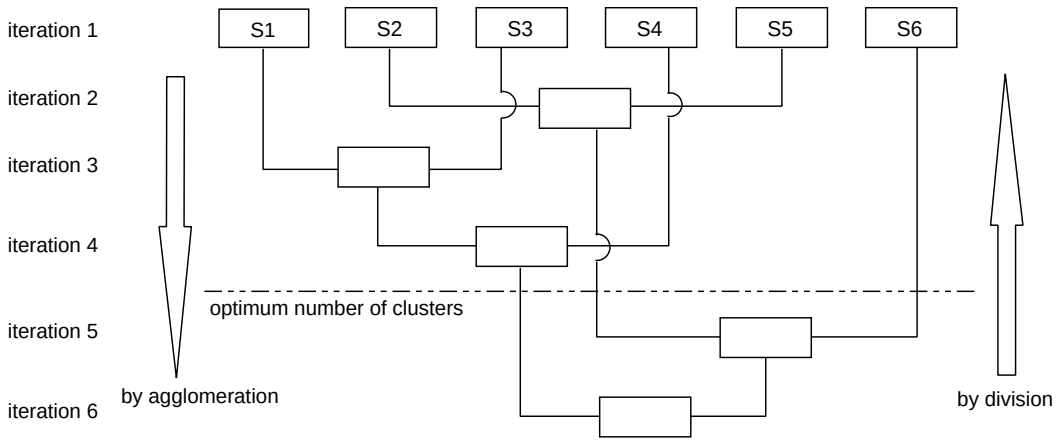


Figure 2.11: Example of the hierarchical clustering techniques.

tering technique begins with all speech segments in a single cluster and iteratively splits clusters to produce new clusters until achieving stopping criterion.

Two common important factors exist in both agglomerative and divisive clustering approaches: distance measure and stopping criterion. A distance measure reflects the dissimilarity between clusters and smaller values indicate that two compared elements are from the same speaker. A stopping criterion is also critical to good clustering performance, as the appropriate number of clusters needs to be decided by terminating the merging or splitting iterations according to the stopping criterion. This stopping criterion may be defined by either thresholding the distance measures or optimizing a certain of model selection criterion.

The distance metrics employed in the speaker change detection (c.f. Section 2.3.2) can also be used for the speaker clustering. However, as each cluster usually consists of more than one segment except for initial clusters in the agglomerative clustering, some methods are required to build the distance between clusters. There are two types of techniques that have been proposed to form the cluster distance. The commonly used approach is to regard each cluster as a single segment, i.e. concatenating all the segments within each cluster into one large segment, and compute the distance between these big segments. Another kind of methods is to construct the distance between two clusters from the distances between the segments within each cluster. According to [Solomonoff *et al.*, 1998], this kind of cluster distance can de defined in different ways:

- **Minimum linkage**: take the smallest distance between a segment from a cluster and one from the other cluster.
- **Maximum linkage**: take the largest distance between component segments from each cluster.
- **Average linkage**: take the averaged value of the distances between all pairs of segments.

All of these three cluster distances have non-distance-like properties. The minimum linkage algorithm gives a small distance to the clusters having a close pair of segment no matter how different the clusters are indeed, whereas both the maximum and average linkage technique produce a non-zero distance value for a cluster and itself, except for the cluster consisting of only one segment. The experiments reported in [Solomonoff *et al.*, 1998] show that the maximum and average linkage algorithm outperform than the minimum linkage for speaker clustering.

B.1 Agglomerative clustering

This is probably the most commonly used speaker clustering technique in the framework of diarization system due to its facility of taking speaker segmentation results as input. Instead of computing a distance value between two adjacent windows in the standard implementation of speaker change detection, a distance matrix is usually used in the agglomerative clustering to give the distances between each pair of clusters. At the beginning of the agglomerative clustering, each speech segment resulting from the speaker change detection stage is created as an initial cluster. At each clustering iteration, the closest pair of clusters are merged into one cluster with the distance matrix being updated. This procedure terminates when a stopping criterion is met. Some representative agglomerative speaker clustering systems using different distance metrics and stopping criteria are introduced in the following.

In [Gish *et al.*, 1991], the GLR distance measure was first applied to the speaker clustering for a particular speech recognition application where an air traffic controller and several pilots are involved in the dialog. After the audio stream is divided into speech segments from individuals by an energy-based speaker segmentation, the utterances from the controller are required to separate from those of the pilots prior to the recognition process. To do this, the authors of [Gish *et al.*, 1991] propose an agglomerative speaker clustering relying on the GLR distance metric, which merges the closest segments into one cluster at the first iteration and repeats the merging process with the cluster distance defined as the maximum GLR from all between-cluster pairs of segments (i.e. maximum linkage). Since two clusters (i.e. one for the controller and the other for pilots) are assumed in this specific application, the clustering algorithm stops when two clusters are reached, with the larger cluster associated with the controller.

The speaker clustering algorithm proposed in [Jin *et al.*, 1997] employs the GLR metric to form the distance matrix, with a weight imposed to the distance between consecutive segments. The agglomerative clustering process stops when the within-cluster dispersion penalized by the number of clusters achieves its minimization:

$$S = |\sum_{i=1}^{K} N_{c_i} \Sigma_{c_i}| * \sqrt{K} \tag{2.47}$$

where K is the number of candidate clusters, N_{c_i} is the number of feature vectors in cluster c_i and Σ_{c_i} is the covariance matrix of cluster c_i, $|.|$ denotes the determinant.

The GLR distance is computed with each segment modeled by a single multivariate Gaussian in [Gish *et al.*, 1991; Jin *et al.*, 1997], while more complex models can be also used when the amount of data within segments is larger. In the work of [Solomonoff *et al.*, 1998] where a set of speech segments each with the duration of 1 minute is given to the speaker clustering, a GMM consisting of 128 diagonal covariance Gaussians is trained for each segment using EM algorithm and then the GLR and a Cross Entropy (CE) distances are used in an agglomerative clustering framework. Given two clusters c_i and c_j, the CE distance is defined as:

$$CE(c_i, c_j) = \log \frac{L(c_i, M_i)}{L(c_i, M_j)} + \log \frac{L(c_j, M_j)}{L(c_j, M_i)} \tag{2.48}$$

where M_i is trained using the data within the cluster c_i. For the stopping criterion, this clustering system adopts the maximization of a cluster purity score based on the distances to nearest neighbor segments. For each cluster c_i, its purity is obtained by averaging the purities of all segments within this cluster:

$$p_{c_i} = \frac{1}{n_{c_i}} \sum_{k=1}^{n_{c_i}} \rho_k \tag{2.49}$$

where n_{c_i} is the number of segments in the cluster c_i and ρ_k is the purity of segment k with respect to the cluster c_i. To compute the segment purity, the distances from one segment to all the other segments are sorted in increasing order and the number of segments that belong to the cluster c_i is counted from the nearest n_{c_i} segments (denoted as n_k^{clus}). The segment purity ρ_k is thus expressed as:

$$\rho_k = \frac{n_k^{clus}}{n_{c_i}} \tag{2.50}$$

The KL2 distance is also successfully used in the context of speaker clustering. The initial speaker clustering implementation using the KL2 distance metric is described in [Siegler *et al.*, 1997], where a given threshold is applied to the distance as the stopping criterion. In this work, the KL2 distance is compared with the Mahalanobis distance and is proven to have better clustering performance. In [Harris *et al.*, 1999], the dissimilarity measure between clusters is given by a variance of the KL2 distance that scales down the distances between neighboring segments by a parameter. It experimentally finds that performing the speaker clustering on different gender/bandwidth conditions separately produces higher cluster purity.

In place of single Gaussian cluster models explored in [Siegler *et al.*, 1997], the agglomerative clustering of [Ben *et al.*, 2004; Moraru *et al.*, 2005] adopts GMM cluster models obtained via the MAP adaptation from the means of the UBM that is trained on the whole speech portions in the incoming audio. This GMM adaptation technique deals with the unreliable parameter estimation problem that is caused by the data insufficiency in clusters at the beginning of the clustering. Relying on the adapted GMM cluster models which have the same component weights

and covariance matrix, a model-space based approximation of the KL2 divergence is proposed as the distance metric between clusters:

$$D(c_i, c_j) = \sqrt{\sum_{m=1}^{N_g} \sum_{n=1}^{d} w_m \frac{(\mu_i(m,n) - \mu_j(m,n))^2}{\sigma_{m,n}^2}} \tag{2.51}$$

where N_g and d are the number of Gaussian mixtures and dimension of acoustic vectors respectively, w_m and $\sigma_{m,n}$ are the common weight and covariance matrix of two GMM cluster models, $\mu_i(m,n)$ and $\mu_j(m,n)$ are the means of two clusters c_i and c_j respectively. A distance threshold stopping criterion is used in the speaker clustering of [Ben *et al.*, 2004], while the BIC criterion is used as the stopping criterion in [Moraru *et al.*, 2005].

In [Reynolds *et al.*, 1998], the dissimilarity between clusters is given by a Cross Likelihood Ratio (CLR):

$$CLR(c_i, c_j) = \log \frac{L(c_i, M_B)}{L(c_i, M_j)} + \log \frac{L(c_j, M_B)}{L(c_j, M_i)} \tag{2.52}$$

where M_B is the Universal Background Model (UBM). The stopping criterion is defined as the maximization of the BBN metric [Solomonoff *et al.*, 1998]:

$$I_{BBN} = \sum_{i=1}^{K} n_{c_i} p_{c_i} - QK \tag{2.53}$$

where K is the number of current clusters, n_{c_i} is the number of segments in the cluster c_i and the cluster purity p_{c_i} is estimated via the nearest neighbor purity algorithm of [Solomonoff *et al.*, 1998] mentioned previously. The parameter Q is designed to control the preference degree of fewer large clusters at the expense of merging more segments from different speakers together.

The speaker clustering method presented in [Nishida and Kawahara, 2003] uses also the CLR distance for a speaker indexing application. As it is well known that GMM is more appropriate for modeling a large and complex data set and VQ modeling method performs better on a small size data set, a BIC-based model selection method is thus applied to determine the optimal cluster model between GMM and VQ according to the amount of data within cluster. The VQ model is derived from a Common Variance GMM (CVGMM) via replacing the covariance matrix with the identity matrix, where CVGMM is constructed by setting each Gaussian mixture with an uniform Gaussian weight and an identical covariance averaged over all GMM covariances of the whole clusters. This speaker clustering system uses a distance threshold stopping criterion.

The predominant stopping criterion used in the field of speaker clustering is the BIC criterion due to its good clustering performance. In [Chen and Gopalakrishnan, 1998a; Chen and Gopalakrishnan, 1998b], the BIC based stopping criterion is originally implemented in conjunction with

an agglomerative clustering framework. This clustering technique assumes a single full covariance Gaussian for each segment and merges iteratively the nearest pair of clusters according to a maximum linkage cluster distance based on the GLR distance, until the minimum ΔBIC (c.f. Equation 2.34) is smaller than a specified threshold (typically 0).

In [Zhou and Hansen, 2000], the clustering system employs a KL2 cluster distance metric with the BIC based stopping criterion and is applied to the segments from the same gender class (male/female). This class dependent clustering method prevents the merging of the segments from different gender speakers and reduces the computational load of pair-wise cluster distances. The authors of [Reynolds and Torres-Carrasquillo, 2004] follows the clustering scheme proposed in [Chen and Gopalakrishnan, 1998b], with the penalty weight λ adjusted to 6.0. In [Moraru *et al.*, 2004b; Moraru *et al.*, 2004c; Meignier *et al.*, 2006], the step-by-step speaker diarization system uses the BIC criterion to estimate the number of speakers. The speaker clustering is carried out individually on each gender/bandwidth combination and the model of each initial cluster is trained using the MAP adaptation of the UBM that is estimated on the whole audio file. The BIC criterion is also used as the distance metric in the agglomerative clustering algorithm of [Moraru *et al.*, 2004b; Moraru *et al.*, 2004c], while the GLR cluster distance is employed in [Meignier *et al.*, 2006].

The speaker clustering implementations of [Chen and Gopalakrishnan, 1998b; Reynolds and Torres-Carrasquillo, 2004; Moraru *et al.*, 2004b; Meignier *et al.*, 2006] computes the penalty term P in the BIC formulation (c.f. Equation 2.34) using the total number of frames in all the clusters, thus the optimal number of clusters is determined globally via the ΔBIC value between K clusters and $K-1$ clusters. Alternatively, a local BIC based stopping criterion may be used to decide whether two particular clusters should to be merged or not, thus the size of the two clusters to be merged is used to calculate the penalty P. This local BIC stopping criterion is applied to an agglomerative clustering scheme in the work of [Tritschler and Gopinath, 1999; Cettolo, 2000] and is shown to perform better than the global one.

The authors of [Moh *et al.*, 2003] propose a novel agglomerative clustering technique that is portable across different data type such as Broadcast News or Conversational Telephone Speech (CTS). Assuming $\{X_k, k = 1, \cdots, S\}$ be the set of segments to cluster, the proposed triangulation method constructs a S-dimensional voice characteristic space by taking the S segments as references and considers each cluster as a point in this space, with its coordinate decided by the distances to all the references. After each segment X_k is modeled by a single full covariance Gaussian M_k, a S-dimensional vector is calculated for each cluster c_i, with the component defined as the conditional probability of the cluster given each segment model (denoted $p(c_i|M_k)$), and then the similarity between clusters is measured by the correlation between the cluster vectors:

$$C(c_i, c_j) = \sum_{k=1}^{S} p(c_i|M_k)p(c_j|M_k) \tag{2.54}$$

where a large correlation value indicates that two clusters are of the same class attribute. This

triangulation algorithm is compared to the standard BIC clustering method on BN and CTS data in [Moh *et al.*, 2003]. It is found that the triangulation technique gives comparable clustering performance compared to the BIC method on BN data, but works less well on CTS data.

Similar to the clustering technique based on the projection of acoustic feature representations into a speaker voice characteristic space in [Moh *et al.*, 2003], the authors of [Tsai *et al.*, 2004] propose to train each segment model via the MAP adaptation of the universal GMM that is estimated using all segments to be clustered. This bottom-up clustering uses a cosine measure as the cluster distance and a thresholding stopping criterion. The references of the speaker voice space are required to reflect the most distinct aspects of voice characteristics. However, since different segments from the same speaker carry similar voice characteristics, the use of each segment as an individual reference tends to warp the speaker voice space. To solve this problem, a improved algorithm based on Eigen Vector Space Model (EVSM) is presented in [Tsai *et al.*, 2005], where a speaker-independent voice space is constructed by applying a Principal Component Analysis (PCA) to reduce the dimensionality of the initial references. The speaker clustering presented in [El Khoury *et al.*, 2007] combines the EVSM based clustering technique with a prosodic based clustering, i.e. a merging of two clusters is performed when their cosine similarity higher is than a given threshold and the difference between the averaged fundamental frequency $F0$ of the two clusters is lower than a threshold.

B.2 Divisive clustering

Divisive clustering is a rarely used speaker clustering technique in the speaker diarization domain. This clustering approach usually starts with one large cluster that is the union of all segments found by a speaker segmentation stage and then splits iteratively clusters, until a stopping criterion is achieved.

One representative implementation of the divisive clustering approach is the speaker clustering system developed at the Cambridge University. This clustering system is originally designed to group together acoustically similar segments so that a Maximum Likelihood Linear Regression (MLLR) speaker adaptation [Leggetter and Woodland, 1995] can be performed robustly to improve speech recognition performance. In [Johnson and Woodland, 1998], a top-down clustering scheme is carried out using two different merging algorithms. Each active node is split into a maximum of 4 child nodes if each child node contains more than a minimum amount of segments. Then, each segment is reassigned to the child node that gives either the minimum covariance-based distance or the maximum likelihood under the model transformed using the data within itself. The covariance matrix or MLLR transform is updated for each child node and this reassignment process is repeated until either no more segments move or a maximum iteration is reached. In this work, both Arithmetic Harmonic Sphericity (AHS) and Gaussian Divergence (GD) distance metrics are employed with each cluster being modeled by a single Gaussian:

$$AHS(c_i, c_j) = \log[tr(\Sigma_{c_j}\Sigma_{c_i}^{-1}) * tr(\Sigma_{c_i}\Sigma_{c_j}^{-1})] - 2\log d \quad (2.55)$$

$$GD(c_i, c_j) = \frac{1}{2}tr(\Sigma_{c_i}^{-1}\Sigma_{c_j} + \Sigma_{c_j}^{-1}\Sigma_{c_i} - 2I) + \frac{1}{2}(\mu_{c_i} - \mu_{c_j})^T(\Sigma_{c_i}^{-1} + \Sigma_{c_j}^{-1})(\mu_{c_i} - \mu_{c_j}) \quad (2.56)$$

where μ_{c_i} and Σ_{c_i} are the mean and covariance matrix of the cluster c_i respectively, d is the dimensionality of feature vectors and I represents the identity matrix.

The AHS-based divisive clustering described above has been applied for the speaker diarization task with some necessary modifications in [Johnson, 1999; Tranter *et al.*, 2004; Tranter and Reynolds, 2004]. The speaker clustering system of [Johnson and Woodland, 1998] imposes a minimum occupancy constraint on each final clusters, which boost the subsequent MLLR speaker adaptation performance by providing sufficient adaptation data. However, the objective of the speaker diarization is to generate one pure cluster for each speaker, where the amount of speech is possible to be small for certain speakers. Therefore, the occupancy restriction is not suitable for the speaker diarization and a stopping criterion is required to determine whether a cluster split should take place. In [Johnson, 1999; Tranter *et al.*, 2004; Tranter and Reynolds, 2004], the AHS-based divisive clustering is performed separately for each gender/bandwidth combination condition, with either a local BIC-based or a cost-based stopping criterion. The cost of each cluster is defined as the mean of distances between its segments and itself and the cost gain is computed as the cost difference between the cluster to be split and the clusters resulting from the splitting.

2.3.4 Integrated segmentation and clustering

Most diarization systems rely on a sequential architecture (c.f. Figure 2.1) where the audio data is segmented first and then clustered. A limitation of this method is that errors made in segmentation step is difficult to correct later, but can also degrade the performance of the subsequent clustering step. An alternative approach is to optimize the segmentation and clustering jointly via iterating both steps such as the systems presented in [Gauvain *et al.*, 1998; Meignier *et al.*, 2000; Ajmera and Wooters, 2003].

The integrated segmentation and clustering approach is first introduced in [Gauvain *et al.*, 1998], where an initial segmentation is provided by a Gaussian divergence measure (c.f. Equation 2.45) based acoustic change detection. The iterative algorithm alternates a Viterbi segmentation and an agglomerative clustering until the maximum value of the objective function is reached. This function is defined as the likelihood of the current segments given their associated clusters clusters with penalized by the weighted sum of the number of segments and and number of cluster. More details about this method will be given in the following chapter (c.f. Section 4.1).

Another integrated diarization system based on an evolutive Hidden Markov model (E-HMM) is proposed in [Meignier *et al.*, 2000; Meignier *et al.*, 2001; Meignier, 2002; Meignier *et al.*, 2006]. The audio recording is represented by an ergodic HMM where each state characterize a

speaker and the transitions model the changes between speakers. This iterative algorithm detects speakers (i.e. cluster) one by one, where both the segmentation and the detection of a new speaker are performed at each iteration using the speaker models detected in the previous iteration. The initial segmentation is modeled by a one-state HMM and the whole audio data is assigned to a single speaker S_0. In each iteration i of the diarization process, a 3 seconds subset from the data labeled as S_0 is selected to represent a non-detected speaker S_i by maximizing the likelihood ratio between model S_0 and a UBM, and then a new speaker model S_i is obtained via MAP adaptation of the UBM using the selected data. A new state corresponding to the speaker S_i is added to the previous HMM and the transition probabilities are updated according to some rules. A new segmentation is generated by relabeling the 3 seconds data from S_0 to S_i. Then, both speaker model adaptation and Viterbi segmentation are repeated until no segmentation change is observed between adaptation/segmentation iterations. The process of adding a new speaker is stopped when the likelihood of the diarization solution has reached the maximum value or when no training data is available to model a new speaker.

The diarization system presented in [Ajmera and Wooters, 2003] also uses an integrated segmentation and clustering method based on HMM. The initialization of this system consists in segmenting the data uniformly into K clusters that are set to over-cluster the data (i.e. K is greater than the number of real speakers) and estimating the GMMs for each initial cluster. Then an Expectation-Maximization (EM) training of the HMM and an agglomerative clustering are performed iteratively until no cluster merging is possible. The HMM training is done in two steps: the first step is to segment the data with the goal of maximizing the likelihood of the data given the cluster GMMs and the second step is to reestimate the parameters of the cluster GMMs from the new segmentation obtained in the previous step. Following each HMM training stage, the two closest clusters are merged into one cluster based on a modified BIC clustering approach. This improved BIC clustering eliminates the penalty term existing in the BIC criterion (c.f. Equation 2.34) by modeling the merged cluster with the number of Gaussian components equal to the sum of those in the two merging cluster candidates.

Chapter 3

Performance metrics and evaluations

This chapter describes the metrics that will be used to score diarization results presented in the following chapters. The first section of this chapter presents the primary performance metric, referred to as overall speaker Diarization Error Rate (DER), that was proposed in the NIST Rich Transcription evaluations. In addition to the overall DER, some metrics defined specially for evaluating the clustering performance are also introduced in this chapter. Finally, an introduction of the main speaker diarization evaluations (i.e. the NIST Rich Transcription evaluations and the French Technolangue ESTER evaluation) is given at the end of this chapter.

3.1 Speaker diarization performance measures

The primary metric for speaker diarization task is the overall speaker Diarization Error Rate (DER) used in the framework of the Rich Transcription (RT) speaker diarization evaluations [NIST, 2003; NIST, 2004; NIST, 2006]. In addition, different evaluation metrics for each stage of speaker diarization systems have also been proposed in the literature. For example, [Delacourt, 2000; Liu and Kubala, 1999] presented the scoring rules for measuring speaker segmentation performance, while [Gauvain *et al.*, 1998; Solomonoff *et al.*, 1998; Chen and Gopalakrishnan, 1998b; Harris *et al.*, 1999] proposed different clustering measurements. Only the overall DER defined by NIST and the clustering metrics proposed by Gauvain *et al.* will be introduced in this thesis and used to present experimental results.

3.1.1 Overall speaker diarization error rate (DER)

The output of diarization systems is a set of segments, each segment composed of a hypothesis speaker ID (i.e. cluster label) with the associated start and end time. The speaker diarization performance can be evaluated by scoring this system output against a reference speaker partition. As described in [NIST, 2003; NIST, 2004; NIST, 2006], the performance is measured by

first computing an optimum one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs for each audio file independently. This mapping is determined so as to maximizing the aggregation over all reference speakers of time that is jointly attributed to both the reference speaker and the mapped (corresponding) hypothesis speaker.

Given the optimum mapping, the overall speaker Diarization Error Rate (DER) is calculated over an entire audio file including overlap regions where multiple reference speakers are talking. This DER score is formulated as:

$$\text{DER} = \frac{\sum_s dur(s) \cdot (\max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))}{\sum_s dur(s) \cdot N_{ref}(s)} \tag{3.1}$$

where the audio is divided into contiguous segments at all speaker change points in which either a reference speaker or a hypothesis speaker starts or stops speaking. Therefore, each s in Equation 3.1 is a section of audio in which the labels (including speaker or non-speech labels) in reference and hypothesis do not change, $dur(s)$ is the duration of s, $N_{ref}(s)$ is the number of reference speakers in s, $N_{hyp}(s)$ is the number of hypothesis speakers in s, $N_{correct}(s)$ is the number of reference speakers for whom their matching hypothesis speakers are also in s.

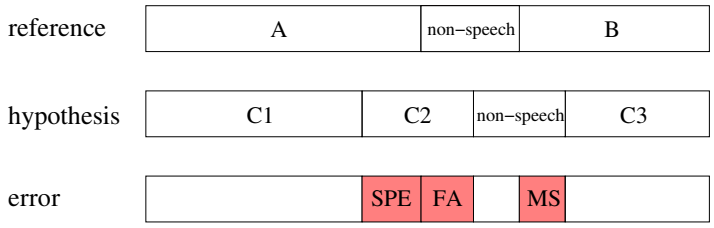


Figure 3.1: Illustration of the overall speaker diarization error rate (DER) used in the NIST Rich Transcription diarization evaluations.

The overall diarization score can be decomposed into three errors coming from the missed speaker, false alarm speaker and speaker match error (c.f. an illustration of all metrics in Figure 3.1). These three error rates are described as follows:

- **Missed speaker error (MS)**: the fraction of speech time that is attributed to more reference speakers than hypothesis speakers. It can be calculated as:

$$\text{MS} = \frac{\sum_s dur(s) \cdot (N_{ref}(s) - N_{hyp}(s))}{\sum_s dur(s) \cdot N_{ref}(s)} \;\; \forall \;\; N_{ref}(s) - N_{hyp}(s) > 0 \tag{3.2}$$

- **False alarm speaker error (FA)**: the fraction of speech time that is attributed to more hypothesis speakers than reference speakers. It can be expressed as:

$$\text{FA} = \frac{\sum_s dur(s) \cdot (N_{hyp}(s) - N_{ref}(s))}{\sum_s dur(s) \cdot N_{ref}(s)} \;\; \forall \;\; N_{hyp}(s) - N_{ref}(s) > 0 \tag{3.3}$$

- **Speaker match error (SPE)**: the fraction of speaker time that is not attributed to a correct speaker (i.e. the mapped reference speaker is not the same as the hypothesis speaker). It can be computed as:

$$\text{SPE} = \frac{\sum_s dur(s) \cdot (\min(N_{hyp}(s), N_{ref}(s)) - N_{correct}(s))}{\sum_s dur(s) \cdot N_{ref}(s)} \tag{3.4}$$

Although the time-based DER can be calculated over a whole audio file, the primary metric employed in RT broadcast news speaker diarization evaluations excludes regions of overlapping speech. Also, a time collar of 0.25 seconds is allowed at each speaker change in the reference in order to take into account possible errors in the reference timing due to the automatic forced-alignment.

It should be noted that this metric is time-weighted, so the DER is biased towards the diarization systems which can correctly detect the main speakers with a large quantity of speech data. The other problem of this measurement is related to the scoring over overlapping segments. As the duration of segments is attributed to all reference speakers who are talking in a segment, this leads to counting the error time more than once. However, the reference data allows to attribute each reference word to its actual speaker, thus the time of each word occurring in the regions of overlap will be counted only once.

3.1.2 Clustering metrics

In order to more closely analyze the performance of speaker clustering methods, average frame-level cluster purity and coverage are used as defined in [Gauvain *et al.*, 1998]. For a given audio stream, let K_s be the number of reference speakers and K_c be the number of clusters generated by diarization systems. Each cluster is a set of segments consisting of a same cluster label and the corresponding start and end time. Let n_{ij} be the number of frames in cluster i which came from speaker j. Likewise, let $n_{i\bullet}$ be the total number of frames in cluster i and $n_{\bullet j}$ be the total number of frames spoken by speaker j. The clustering performance can be evaluated by two types of scoring:

- **Cluster purity**: for a cluster i, the purity $p(i)$ is defined as the ratio between the number of frames by the dominating speaker in the given cluster and the total number of frames in this cluster:

$$p(i) = \max_j \frac{n_{ij}}{n_{i\bullet}} \tag{3.5}$$

This metric form was used by [Chen and Gopalakrishnan, 1998b] at a segment level, then it was employed by [Gauvain *et al.*, 1998; Harris *et al.*, 1999] at a frame level. The mean purity will be used to assess clustering performance in this thesis, which is obtained by time-weighted averaging over all cluster purities:

$$purity = \frac{\sum_{i=1}^{K_c} n_{i\bullet} p(i)}{\sum_{i=1}^{K_c} n_{i\bullet}} \tag{3.6}$$

- **Cluster coverage**: for a reference speaker j, the cluster coverage $q(j)$ is defined as the percentage of its frames in the cluster which has most of the data of this speaker. This is a measure of the dispersion of a given speaker's data across clusters:

$$q(j) = \max_i \frac{n_{ij}}{n_{\bullet j}} \tag{3.7}$$

This metric has also been referred to as the purity of reference clusters in [Cettolo, 2000]. The mean coverage is given as

$$coverage = \frac{\sum_{j=1}^{K_s} n_{\bullet j} q(j)}{\sum_{j=1}^{K_s} n_{\bullet j}} \tag{3.8}$$

The cluster purity and coverage evaluate complementary aspects of the problem of clustering. In fact, optimizing the former tends to produce small clusters, this is helpful to reduce the risk of grouping segments from different speakers into a same cluster. However, a cluster purity of 100% can be obtained from a clustering with a cluster per frame. On the contrary, optimizing the latter tends to produce large clusters, this is favorable to avoid classifying segments from the same speaker into different clusters. However, the maximum coverage can be derived from a clustering algorithm that groups all frames into a single cluster. Hence, the two scores allow to tune clustering methods according to the specific requirements of the problem under consideration.

The speaker match error presented previously can be interpreted as a combination of cluster purity and coverage errors. Moreover, it is interesting to note that if, the majority reference speaker ID in a cluster is always mapped into the hypothesis speaker ID for this cluster, then the speaker match error will be exactly the cluster purity error; it is easy to demonstrate that it is also a lower bound for the match error. Thus, starting with an initial segmentation, the cluster purity error will be the lowest possible match error on this segmentation after performing an agglomerative clustering. The same holds for further clustering of the output of a previous clustering stage, as long as the segment boundaries are not modified.

3.2 Speaker diarization evaluations

During the work of this thesis, the proposed diarization systems have been used to participate to the different year's diarization evaluations organized by NIST and a French evaluation campaign namely ESTER. The following is a summary introduction of these evaluations.

3.2.1 NIST evaluation campaigns

The National Institute of Standards and Technology (NIST) is a non-regulatory federal agency of the U.S. Commerce Department's Technology Administration, whose task is to advance measurement science, standards and technology so as to promote U.S. innovation and industry. In order to exploit spoken language as an alternative modality for the human-computer interface, the speech group within NIST encourages the advancement of the spoken language processing technologies (including speech recognition and understanding) by developing measurement methods, proving reference materials, organizing benchmark tests within the research and development community and building prototype systems.

NIST has been coordinating a series of yearly evaluations in the topic of speaker recognition since 1996 [NISTSRE] for the purpose of indicating the direction of research efforts and providing a calibration of technical capabilities. These evaluations focus on the automatic speaker detection task and supported the speaker segmentation task from 2000 to 2002. The term "speaker segmentation" used within the framework of Speaker Recognition Evaluations (SRE) had a same definition as the speaker diarization that was introduced in the series of Rich Transcription evaluations in 2003. Both terms represent the task consisting of segmenting an audio stream into speaker turns and grouping all turns spoken by the same speaker. From SRE 2000 to 2001, the data used for the speaker segmentation task were summed two-channel telephone conversations and included two different types of data. The first type consisted of segments with a duration of approximately one minute, which were known to contain exactly two speakers. The second type was composed of segments of different duration, up to a maximum of 10 minutes. The number of speakers in each of these segments was unknown a priori and may be as many as ten in a segment. In SRE 2002, the data were drawn from a variety of different sources, including telephone conversations, broadcast news and meetings and the number of speakers in each segment varied and was not given in advance.

The Rich Transcription (RT) evaluation series conducted by NIST was started in 2002 until the latest one in 2007 [NISTRT]. The objective of RT evaluations is to create speech recognition technologies that will make transcriptions more readable by humans and more useful for machines. This is done by integrating a transcription of the spoken words provided by a Speech-to-text (STT) system with metadata information about speakers, discourse (e.g. disfluencies, emotion) and much more. Therefore, the speaker diarization task has been included in Rich Transcription evaluations as a sub-task of the Metadata Extraction (MDE). The speaker diarization task was evaluated for broadcast news (BN) and conversational telephone speech (CTS) in English from 2002 to 2004, then its focus has been changed to meeting domain since 2005. The data type used

for speaker diarization task in each RT evaluation is overviewed as follows:

- Rich Transcription 2002 Evaluation **(RT-02)**: English broadcast news (BN) data recorded on single audio channel and English conversational telephone speech (CTS) data recorded on two distinct audio channels (one for each conversation side)
- Rich Transcription Spring 2003 Evaluation **(RT-03S)**: English BN and CTS data
- Rich Transcription Fall 2003 Evaluation **(RT-03F)**: English BN and CTS data
- Rich Transcription Spring 2004 Meeting Recognition Evaluation **(RT-04S)**: English meeting recordings in Multiple Distant Microphones (MDM) and Single Distant Microphone (SDM) conditions
- Rich Transcription Fall 2004 Evaluation **(RT-04F)**: only English BN data
- Rich Transcription Spring 2005 Meeting Recognition Evaluation **(RT-05S)**: conference room and lecture room meetings in English on MDM, SDM and three different microphone arrays conditions
- Rich Transcription Spring 2006 Meeting Recognition Evaluation **(RT-06S)**: conference room and lecture room meetings in English on MDM, SDM and two different microphone arrays conditions
- Rich Transcription Spring 2007 Meeting Recognition Evaluation **(RT-07S)**: conference room and lecture room meetings in English on MDM, SDM, ADM (All Distant Microphones) and two different microphone arrays conditions

3.2.2 ESTER evaluation campaign

The ESTER evaluation campaign [Gravier *et al.*, 2004] is a part of the EVALDA project dedicated to the evaluation of language technologies for the French language and organized jointly by the Francophone Speech Communication Association (AFCP), the French Defense expertise and test center for speech and language processing (DGA/CEP) and the Evaluation and Language resources Distribution Agency (ELDA). This campaign started in 2003 and consisted of a Phase I dry run in January 2004 and a Phase II official test in January 2005. The aim of the ESTER evaluation is to evaluate broadcast news rich transcription systems for the French language, which complements the NIST RT evaluations on the English, Arabic and Chinese languages. It focus on three categories of tasks: orthographic transcription, segmentation (including sound event detection, speaker tracking and speaker diarization tasks) and named entity detection. The datasets used in ESTER were drawn from multiple France radio broadcast sources.

Chapter 4

Speaker diarization for Broadcast News

The descriptions of the baseline and improved speaker diarization systems are given in this chapter. The baseline speaker partitioning system was developed for the LIMSI broadcast news transcription system, providing a high cluster purity, along with a tendency to split data from speakers with a large quantity of data into several segment clusters. Based on this partitioner, a multi-stage diarization system integrates a Bayesian Information Criterion (BIC) agglomerative clustering to replace the iterative Gaussian mixture model (GMM) clustering, and then it combines a second clustering stage relying on a GMM-based speaker identification method. A final post-processing stage refines the segment boundaries using the output of a transcription system. This improved multi-stage system provided the state-of-the-art diarization performance in the RT-04F and ESTER evaluations. The experimental results on both evaluations datasets are presented at the end of this chapter.

4.1 Baseline partitioning system

The baseline audio partitioning system was developed as a preprocessing step for the LIMSI English Broadcast News transcription system [Gauvain *et al.*, 1998; Gauvain *et al.*, 2001]. As the objective of the automatic transcription is the minimum word error rate, the baseline partitioning system is required to provide accurate segment boundaries rather than minimum diarization errors. Although the rejection of non-speech segments is useful in order to minimize insertion of words and to save computation time, it is important that the segment boundaries are located in non-informative zones such as silences or breaths. Indeed, having a word cut by a boundary disturbs the transcription process and increases the word error rate. It has been found that the baseline partitioning system provides a high cluster purity, but has a tendency to split speech from speakers with a large quantity of data into several segment clusters that may correspond to different acoustic environments. This section will focus into introducing this baseline partitioner **c-std** shown in Figure 4.1.

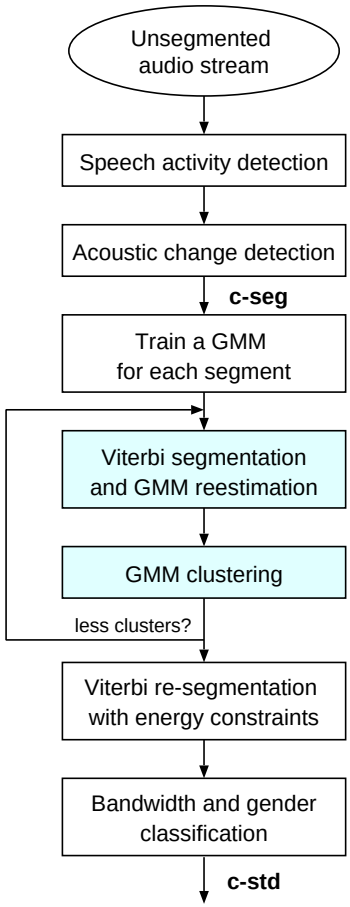


Figure 4.1: Block diagram of the baseline audio partitioning system.

4.1.1 Feature extraction

Mel frequency cepstral parameters are extracted from the speech signal every 10 ms using a 30 ms window on a 0-8kHz band. For each frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform.

Using a process similar to that of PLP computation [Hermansky, 1990], 12 LPC-based cepstral coefficients are then extracted. The 38 dimensional feature vector consists of 12 cepstral coefficients, Δ and Δ-Δ coefficients plus the Δ and Δ-Δ log-energy. This is essentially the same set of features that is used in a standard transcription system, except for the energy [Gauvain *et al.*, 2002].

This set of acoustic parameters is used in all steps of the **c-std** system, except for the acoustic change detection where only the static features are used. No cepstral mean or variance normalization is performed to the acoustic vector in the baseline partitioning system.

4.1.2 Speech Activity Detection (SAD)

To detect speech segments, a Viterbi decoder is performed on the input audio with using Gaussian mixture models (GMMs). These GMMs, each with 64 Gaussians, represent individually speech, noisy speech, speech over music, pure music, silence and advert. This speech activity detector was designed to remove only long regions without speech such as silence, music and noise, so the penalty of switching between models in the Viterbi decoding was set to minimize the loss of speech signal. The GMMs were each trained on about 1 hour of acoustic data, selected from English Broadcast News data from 1996 and 1997 distributed by the Linguistic Data Consortium (LDC). The specific training data for each acoustic model are detailed as follows:

- Speech: trained on data of all types, with the exception of pure music segments and silence portions from segments transcribed as speech in the presence of music or background noise
- Noisy speech: trained on speech occurring with noise
- Speech over music: trained on speech occurring with music
- Pure music: trained on pure music segments
- Silence: trained on silence segments except those in segments labeled as speech in the presence of music or background noise
- Advert: trained on segments which were labeled as speech over music and occur repeatedly in BN shows

For generating the final SAD output, all segments labeled as 'noisy speech', 'speech over music' or 'advert' are first reclassified as speech segments, then the contiguous speech segments are merged into a segment. Only speech segments will be passed to the clustering stage.

4.1.3 Acoustic change detection

The acoustic change detection is carried out on the whole audio stream to locate segment boundaries that correspond to the instantaneous acoustic change points. As discussed in Section 2.3.2, acoustic change detection is an extension of speaker change detection, it detects both speaker changes and other acoustic changes between different audio sources such as speech, music, silence or background noise. Therefore, the number of segments produced by this step is much larger than the true speaker turns. Such segmentation provides a relatively low miss rate of speaker change points but a high false alarm rate, while the false change points can be easily removed later during a clustering process.

This acoustic change detection is a kind of metric-based segmentation and is performed by taking the maxima of a local Gaussian divergence measure between two adjacent sliding windows s_1 and s_2. For each segment, the static features (i.e., only the 12 cepstral coefficients plus the

energy) are modeled with a single diagonal Gaussian, i.e. $s_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $s_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ with Σ_1 and Σ_2 diagonal. Then the Gaussian divergence measure is defined as:

$$G(s_1, s_2) = (\mu_2 - \mu_1)^T \Sigma_1^{-1/2} \Sigma_2^{-1/2} (\mu_2 - \mu_1) \tag{4.1}$$

It is the Mahalanobis distance between μ_1 and μ_2 weighted by the geometric mean of Σ_1 and Σ_2, which reduces to a weighted Euclidean distance because of the diagonal assumption. The detection threshold was optimized on the training data in order to provide acoustically homogeneous segments. The window size was set to 5 seconds with a minimal segment length of 2.5 seconds. Due to the simple diagonal assumption, this acoustic change detection phase is very quick. This approach is similar to that proposed in [Siegler *et al.*, 1997] using the symmetric KL2 metric. Other popular detection methods are based upon the BIC metric [Chen and Gopalakrishnan, 1998b; Delacourt and Wellekens, 2000; Cettolo, 2000], but these methods show a much higher complexity [Cettolo *et al.*, 2005]. The various acoustic change detection techniques based on energy, models or metrics have been introduced in Section 2.3.2.

4.1.4 Iterative GMM segmentation/clustering procedure

Each initial segment is used to seed one cluster, and an 8-component GMM with a diagonal covariance matrix is trained on the segment's data. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments produced by the SAD stage. This iterative algorithm alternates the Viterbi resegmentation and the GMM re-estimation and merging steps, with the goal of maximizing the objective function:

$$\sum_{i=1}^{N} \log f(s_i | M_{c_i}) - \alpha N - \beta K \tag{4.2}$$

where $S = (s_1, \cdots, s_N)$ is the partitioning of the speech segments into a sequence of N segments, $c_i \in [1, K]$ is the cluster label for the segment s_i among the K different clusters, $f(s_i|M_{c_i})$ is the likelihood of the segment s_i given the model of its cluster M_{c_i}, and α and β are the segment and cluster penalties. The procedure stops when no more likelihood increase is obtained after the merge of clusters. More details on the clustering procedure can be found in [Gauvain *et al.*, 1998]. This procedure is similar to BIC using a global penalty as described in the next section.

4.1.5 Viterbi re-segmentation

The segment boundaries are refined using the last set of GMMs and an additional relative energy-based boundary penalty, within a one second interval. The boundaries are thus shifted to the

nearest point of low energy within this interval. This is done so as to locate the segment boundaries at silence portions, thereby avoiding cutting words. This is especially important when using the resulting segmentation as a pre-processing step of an automatic transcription system.

4.1.6 Bandwidth and gender labeling

Bandwidth (wide band studio or narrow band telephone) detection for each segment is first performed using maximum likelihood classification with two speaker-independent GMMs. The gender (male or female) labeling is then carried out on the segments using two pairs of bandwidth dependent GMMs: for gender labeling on telephone speech segments, feature extraction is limited to the 0-3.5kHz band. The GMM models are composed of 64 components with diagonal covariance matrices and were trained on the subset of the LDC 1996/1997 English Broadcast News data also used to train the speech detection models. This labeling is useful for the transcription system, as different acoustic models are used for each combination of bandwidth and gender for better performance, but is also of interest for structuring the acoustic stream by supplying extra information about the speakers in the final output. As it is possible for speech from the same speaker to occur in different parts of the same BN show with different channel conditions (i.e. studio or telephone), performing the bandwidth labeling on a segment level rather than on a whole cluster may split a cluster into different bandwidth classes.

4.2 Multi-stage diarization for broadcast news

As introduced in Section 2.3.3, recent researches have shown BIC clustering methods to obtain good performance on the speaker diarization task [Moh *et al.*, 2003; Tranter and Reynolds, 2006]. The baseline system was therefore modified by replacing the iterative GMM clustering with a Bayesian information criterion (BIC) agglomerative clustering using single Gaussians with full covariance matrices. In addition, since different models can capture various and complementary aspects of the data, a combination of different clusterings is designed to make use of single Gaussians and GMMs. This is done by pipelining the output of the BIC clustering into a second clustering stage using a GMM-based speaker identification (SID) method, where the BIC clustering is optimized to generate a under-clustered output so as to give the SID clustering stage pure clusters with a reasonable amount of speech. This SID clustering employs a more aggressive acoustic channel normalization technique and a more complex speaker model, which is enabled by the larger amount of data per cluster provided by the BIC clustering. Finally, an SAD post-filtering stage is carried out to remove short intra-speaker pauses. These short pauses are indeed harmless for a speech transcription system, while they are penalized as false alarm speech errors for a speaker diarization system.

As shown in Figure 4.2, the multi-stage speaker diarization system is constructed by integrating the modifications as mentioned above with the baseline partitioning system. The new modules employed in the diarization system will be described in the rest of this section.

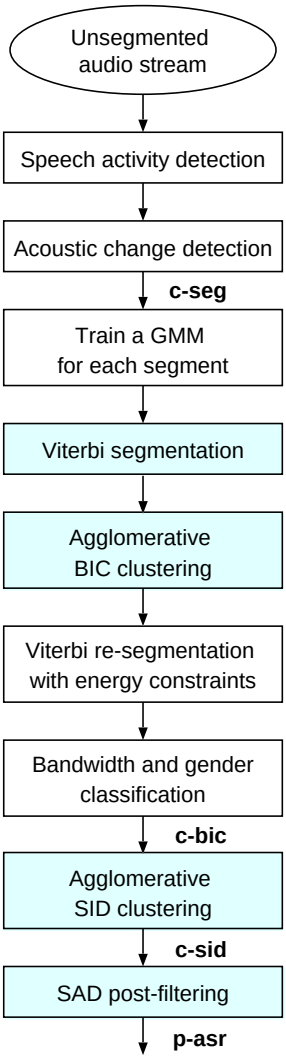


Figure 4.2: Block diagram of the multi-stage speaker diarization system for Broadcast News.

4.2.1 BIC clustering

Agglomerative clustering is applied to the segments resulting from the GMM segmentation. Initially, each segment seeds one cluster, modeled by a single Gaussian with a full covariance matrix trained on the 12 Mel frequency cepstrum coefficients and the energy. Here, the Δ and Δ-Δ coefficients are not used for the purpose of decreasing the computational load caused by the large dimension of full covariance matrices. At each iteration, the two nearest clusters are merged until the stopping criterion is reached. The BIC criterion [Chen and Gopalakrishnan, 1998b] is used both for the inter-cluster distance measure and the stop criterion.

In order to decide whether to merge two clusters c_i and c_j, the ΔBIC value is computed as:

$$\Delta BIC = (n_i + n_j) \log |\Sigma| - n_i \log |\Sigma_i| - n_j \log |\Sigma_j| - \lambda P \tag{4.3}$$

where Σ is the covariance matrix of the merged cluster (c_i and c_j), Σ_i of cluster c_i, Σ_j of cluster c_j, and n_i and n_j are respectively the number of the acoustic frames in clusters c_i and c_j. The penalty P is:

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log n \tag{4.4}$$

where d is the dimension of the feature vector space. The term $n_i \log |\Sigma_i|$ is related to the log likelihood of the cluster c_i given its estimated Gaussian M_{c_i}[1]. Singular covariance matrices were not an issue because of the minimal length constraint during the acoustic change detection. The merging criterion is that two clusters should be merged if $\Delta BIC < 0$. At each step the two nearest clusters (i.e. those which have the most negative ΔBIC value) are merged into one cluster, and the ΔBIC values between this new cluster and remaining clusters are computed. This clustering procedure terminates when the ΔBIC between all cluster pairs is greater than zero.

In our BIC clustering procedure, the size of the two merged clusters, i.e. $n = n_i + n_j$, is used to compute the penalty P, as described in [Cettolo, 2000]. We refer to this as a local BIC penalty. Another solution is to use the size of the whole set of clusters, i.e. $n = \sum_{k=1}^{N} n_k$ to compute the penalty, which we refer to as a global BIC penalty and corresponds to an exponential prior for the number of clusters. In this case the penalty is constant and the decision to merge two clusters is decided just by the increase in likelihood. This in fact corresponds to the objective function in Equation (4.2) used in the baseline partitioner when the number of segments is fixed. For Broadcast News documents, the experimental results as presented in Section 4.3 demonstrate the local BIC to be a better choice for a merging criterion.

4.2.2 SID clustering

After the initial Viterbi segmentation, both the iterative GMM and the agglomerative BIC clustering methods have to deal in the beginning of the process with short duration segments, and thus use a limited set of parameters per cluster. After several iterations of clustering, the amount of data per cluster increases, so a more complex model can be used. Our approach is to stop the initial clustering stage early, and use the results to seed a second clustering stage with more initial data per cluster. This second stage can therefore estimate more complex models for the speakers. In addition, purely acoustic clustering tends to split a speaker's data into several clusters as a function of the various background conditions (clean speech, speech with noise, speech

[1] more precisely, $\log f(c_i|M_{c_i}) = -\frac{n_i}{2} \log |\Sigma_i| - \frac{n_i d}{2}(1 + \log 2\pi)$, but the constant factor $\frac{1}{2}$ in the term $\log |\Sigma_i|$ was simplified in Equation 4.3.

with music etc.), so an acoustic background normalization is necessary to regroup the data for a given speaker.

After the BIC clustering stage, state-of-the-art speaker recognition methods [Schroeder and Campbell, 2000; Barras and Gauvain, 2003] were used to improve the quality of the speaker clustering.

The feature vectors employed in this SID clustering stage consists of 15 Mel frequency cepstral coefficients plus delta coefficients and delta energy. These acoustic parameters differ from those used for the other modules in the multi-stage diarization system (c.f. Section 4.1.1). As noted in [Barras and Gauvain, 2003], better speaker recognition performances can be obtained by using 15 cepstral coefficients rather than 12, since higher order cepstral coefficients are known to carry some extra speaker information. It has also been experimentally found that the second-order derivatives are not essential for speaker recognition [Doddington *et al.*, 2000]. After that, feature warping normalization as presented in Section 2.2.2, which reshapes the histogram of the cepstral coefficients into a Gaussian distribution [Pelecanos and Sridharan, 2001] is performed on each segment using a sliding window of 3 seconds in order to reduce the effect of the acoustic environment.

For each gender and channel condition (studio, telephone) combination, a Universal Background Model (UBM [Reynolds *et al.*, 2000]) with 128 diagonal Gaussians is trained on the 1996/1997 English Broadcast News data. The GMM for each remaining cluster is obtained by maximum a posteriori (MAP) adaptation [Gauvain and Lee, 1994] of the means of the matching UBM. As demonstrated in Section 2.2.3, the adapted mean parameters are calculated using the following formula:

$$\mu_i = \frac{c_i E_i(x) + r\mu_{i,ubm}}{c_i + r} = \frac{\sum_{n=1}^{n_k} p(i|x_n, \Theta_{ubm}) x_n + r\mu_{i,ubm}}{\sum_{n=1}^{n_k} p(i|x_n, \Theta_{ubm}) + r} \tag{4.5}$$

where c_i is the *a posteriori* probability of the ith Gaussian component given the speech data $x_1, \cdots, x_{n_k}$ from cluster c_k, r is a predefined relevance factor controlling the balance between the parameters from the prior UBM and the given cluster data.

Agglomerative clustering is performed separately for each gender and bandwidth condition (c.f. Figure 4.3), using a cross log-likelihood ratio as in [Reynolds *et al.*, 1998]. For each cluster c_i, its model M_i is MAP adapted from the gender and channel matched UBM B using the feature vectors x_i belonging to the cluster. Given two clusters c_i and c_j, the cross log-likelihood ratio $\mathcal{S}$ is defined as:

$$\mathcal{S}(c_i, c_j) = \frac{1}{n_i} log \frac{f(x_i|M_j)}{f(x_i|B)} + \frac{1}{n_j} log \frac{f(x_j|M_i)}{f(x_j|B)} \tag{4.6}$$

where $f(\cdot|M)$ is the likelihood of the acoustic frames given the model M, and n_i is the number of frames in cluster c_i. This is a symmetric similarity measure. After each merge, a new model

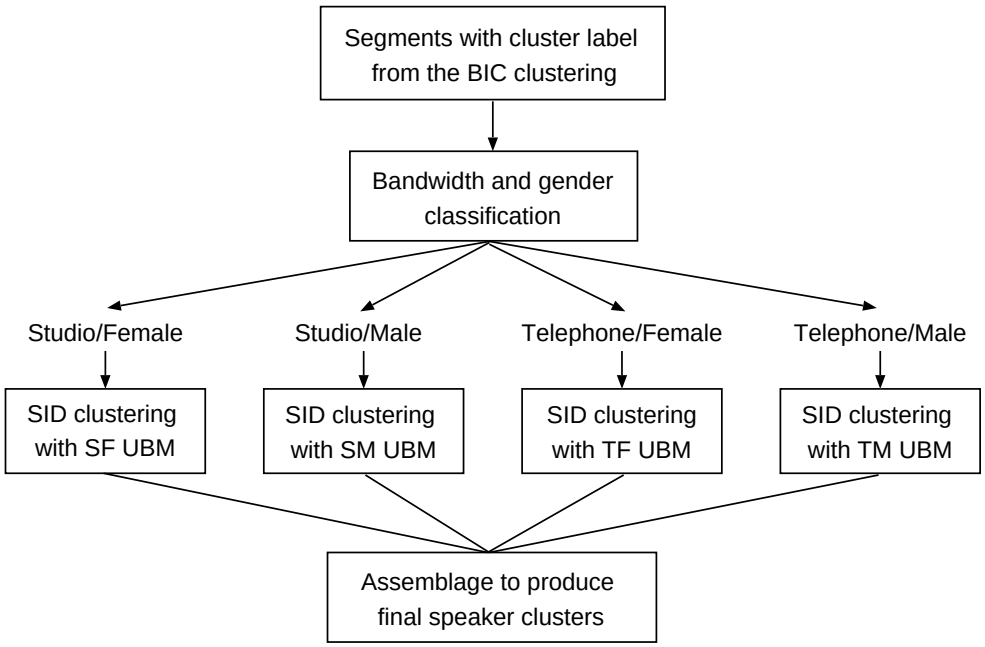


Figure 4.3: SID clustering stage applied on each gender and bandwidth combination.

is trained for the cluster $c_i \cup c_j$. The clustering stops when the cross log-likelihood ratio between all clusters is below a given threshold δ optimized on the development data.

4.2.3 SAD post-filtering

As discussed in Section 4.1.2, the Viterbi decoder with 6 broad acoustic models aims to detect the speech segments, thus only non-speech segments with a sufficient long-duration are found out for minimizing the missed speech error. In order to filter out short-duration silence segments that were not removed in the initial SAD step to further reduce the speaker diarization error, especially the false alarm speech error, a post-processing stage uses the word segmentation output by the LIMSI Broadcast News Speech-To-Text system [Nguyen *et al.*, 2004] relying on the **c-std** system for the segmentation and clustering. Only inter-word silences longer than 1 second are filtered out, this value being determined empirically.

4.3 RT-04F experiments

This section reports diarization experiments conducted on the US English Broadcast News databases from the Rich Transcription Fall 2004 (RT-04F) evaluation [NIST, 2004]. The experiments were performed using both the baseline speaker diarization system described in Section 4.1 and the

proposed multi-stage diarization system described in Section 4.2. In 2004, LIMSI participated to the speaker diarization task in the RT-04F evaluation, the results from LIMSI's submitted diarization systems are also given at the end of this section. The metrics presented in Section 3.1 are used to measure speaker diarization performances and the primary metric employed in the RT-04f evaluation is the overall speaker diarization error rate (DER) that is calculated on the whole audio data excluding overlapping speech regions.

4.3.1 Databases description

The training and development datasets were provided for system development along with a reference speaker labeling determined by the Linguistic Data Consortium (LDC). Evaluation references were made available after the evaluation. The data are extracted from multiple US radio or TV Broadcast News shows. The style of shows varied from a set of discourses from a few speakers to rapid headline news reporting. The development data has two portions, *'dev1'* and *'dev2'* each consisting of 6 30-minute audio files. The evaluation data is comprised of 12 audio files each lasting approximately 30 minutes. The *dev2* and evaluation datasets are very similar to each other since the shows are recorded from the same broadcast sources, at the same period, whereas *dev1* is an older corpus coming from the previous RT-03 evaluation. More detailed information about the databases used in RT-04F are shown in Table 4.1.

dataset	*show nb*	*show duration*	*epoch*	*source*
train	23	30 min. - 2 hr.	Jul 1997 -Jan 1998	ABC CNN CSPAN PRI
dev1	6	30 min.	Feb 2001	ABC CNN NBC PRI VOA
dev2	6	30 min.	Nov 2003 -Dec 2003	ABC CNBC CNN CSPAN PBS
eval	12	30 min.	Dec 2003	ABC CNBC CNN CSPAN PBS WBN

Table 4.1: The databases used in the RT-04F evaluation.

4.3.2 System configurations

Unless other specification, the default configuration is the one that provided the best result on a subset of the RT-04F development data during the RT-04F evaluation period, i.e. $\alpha = \beta = 230$ for **c-std**, $\lambda = 5.5$ for **c-bic** and $\lambda = 3.5, \delta = 0.1$ for **c-sid** and **p-asr**. In the **c-sid** system, the BIC penalty weight λ was optimized to cluster only the closest segments in the BIC clustering stage so as to give more degrees of freedom to the SID clustering stage; while in the **c-bic** system, λ was optimized directly to give the lowest diarization error the BIC clustering could bring. A local BIC merging and stopping criterion discussed in Section 4.2.1 was also used.

4.3.3 RT-04F development results

The performance at different stages of the system was compared for different system configurations, as presented in Table 4.2. On RT-04F *dev1* and *dev2*, the initial segmentation **c-seg**, with the minimal duration constraint of 2.5 sec per segment, has a purity error of 1% which is also the lowest possible speaker match error. The coverage error of this initial segmentation is of course very high at 73.2% on *dev1* and 81.2% on *dev2*; which means that on average, only about one quarter of the speech for each speaker is located in a single segment.

As expected, the standard partitioner **c-std** in its default configuration (i.e $\alpha = \beta = 160$) provides good purity, but relatively poor coverage, resulting in a high overall diarization error of 32.3% on *dev1* and 37.3% on *dev2*. Setting the penalty α and β to optimize these values reduces this error around 25%. The **c-bic** system also provides a high purity (2.9% and 3.4% purity error respectively on *dev1* and *dev2*), with a much better coverage (9.8% and 13.5% coverage error respectively on *dev1* and *dev2*), reducing relatively the overall error rate by 47% on *dev1* and 33% on *dev2*. The **c-sid** system achieves a large decrease of the coverage error with a further small improvement of the purity, resulting in an overall DER of 7.1% on *dev1* and 7.6% on *dev2*, a relative reduction of approximate 50% compared with the **c-bic** system.

system	*purity error (%)*		*coverage error (%)*		*overall DER (%)*	
	dev1	*dev2*	*dev1*	*dev2*	*dev1*	*dev2*
c-seg	1.0	1.0	73.2	81.2	N/A	N/A
c-std ($\alpha = \beta = 160$)	5.0	4.1	28.4	31.9	32.3	37.3
c-std ($\alpha = \beta = 230$)	9.4	7.3	17.9	20.6	24.8	27.8
c-bic ($\lambda = 5.5$)	2.9	3.4	9.8	13.5	13.2	18.6
c-sid ($\lambda = 3.5, \delta = 0.1$)	2.1	1.7	4.2	3.5	**7.1**	**7.6**

Table 4.2: The purity, coverage and overall diarization error rates from the **c-std**, **c-bic** and **c-sid** systems on two RT-04F development datasets.

Looking at the performance of the **c-sid** system in more detail, a large variation in the speaker match error is observed across shows, as shown in Table 4.3. The fields of the show name in Table 4.3 are delimited by '_', the first field is the show date (i.e. month plus day), the next is the show duration in minutes, and the last presents the show source. The speaker error ranges from a low of 0.1% to over 12%. Having only very few speakers (3), the "1115_30_CSPAN" show has the lowest speaker error. The "0206_30_ABC", "0221_30_NBC" and "1127_30_ABC" shows have more speakers, occurring in different background conditions, which is more challenging for the diarization system. These facts imply that there is a certain correlation between the diarization performance and the composition of the shows, such as number of speakers or recording conditions, etc.

To further investigate the variability of the system performance over the shows, the speech time per speaker is calculated for each show individually, as demonstrated in Figure 4.4. It can be

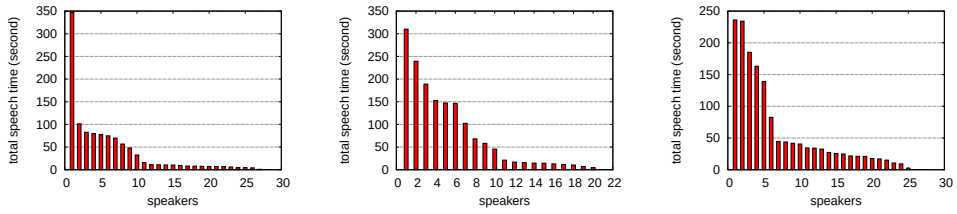
show	*#REF*	*#SYS*	*MS (%)*	*FA (%)*	*SPE (%)*	*DER (%)*
dev1	119	161	**0.4**	**1.3**	**5.4**	**7.1**
0206_30_ABC	27	37	1.6	1.3	12.4	15.2
0217_30_VOA	20	22	0.3	1.2	2.2	3.7
0220_30_PRI	25	30	0.1	0.9	2.8	3.8
0221_30_NBC	21	35	0.1	1.1	12.0	13.2
0225_30_CNN	16	21	0.5	1.4	5.6	7.6
0228_30_MNB	10	16	0.2	1.8	0.8	2.8
dev2	90	130	**0.5**	**3.1**	**4.1**	**7.6**
1115_30_CSPAN	3	4	0.3	2.9	0.1	3.3
1118_30_CNN	17	22	0.6	4.2	5.0	9.8
1120_30_PBS	27	29	0.1	2.8	7.4	10.3
1127_30_ABC	23	29	2.1	6.7	12.5	21.2
1129_30_CNNHL	9	26	0.0	1.4	0.5	1.9
1201_30_CNBC	11	20	0.2	1.0	0.9	2.1

Table 4.3: Per-show and total diarization results on two RT-04f development databases from the **c-sid** system, #REF and #SYS are respectively the reference and system speaker number.

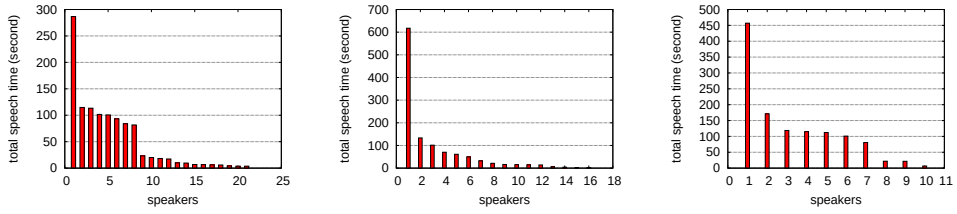
found that the shows with the speaker number less or around 10 have very low speaker match errors, generally less than 1% (e.g. “0228_30_MNB”, “1115_30_CSPAN”, “1129_30_CNNHL” and “ 1201_30_CNBC”). This indicates that the data with less speakers is normally easier for the speaker diarization task. For the shows with more speakers, the speaker match error keeps relatively low when the distribution of the speech time for each speaker is relatively smoothly (c.f. sub-figure (b) and (c)). The speaker match error increases largely with being higher than 5%, when there is one speaker who spoke much more than the other speakers in a show (c.f. sub-figure (e) and (h)). In the same case, a more rapid decrease of diarization performance is observed when there are more speakers in data (c.f. sub-figure (a), (d) and (i)). However the show “1127_30_ABC” is an exception, it has a high speaker error rate with a sooth speaker time distribution.

4.3.4 Local vs. global BIC on RT-04F development data

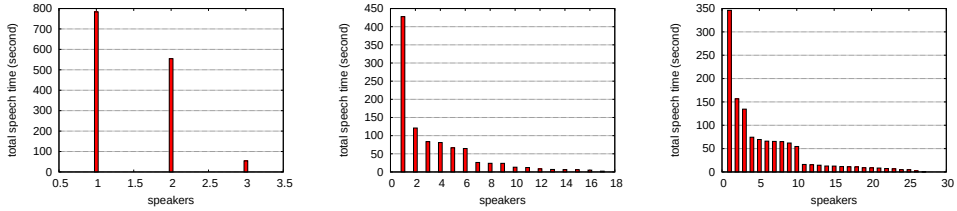
In order to compare the clustering performance from the local and global BIC criteria, some post-evaluation experiments were ran on the entire RT-04F development data (i.e. both *dev1* and *dev2*). As described in Section 4.2.1, the BIC criterion is used both as the inter-cluster distance measure and the clustering stopping criteria. The results from changing the λ penalty using both the local and global BIC implementation in the **c-bic** and **c-sid** systems are illustrated in Figure 4.5. The DER curves of the local BIC implementation show that there are no significant DER changes between the optimal λ value and the selected value for the RT-04F evaluation (1.3% DER difference for the **c-bic** system and only 0.2% for the **c-sid** system). The local BIC



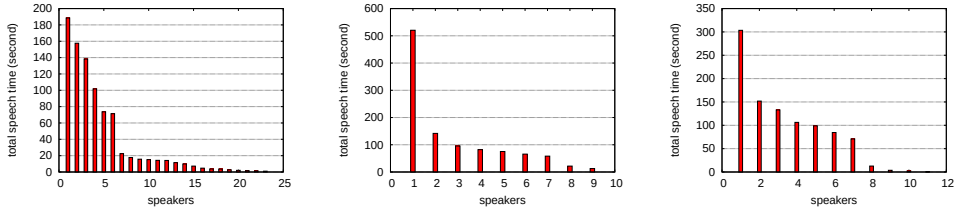
(a) 0206_30_ABC (b) 0217_30_VOA (c) 0220_30_PRI



(d) 0221_30_NBC (e) 0225_30_CNN (f) 0228_30_MNB



(g) 1115_30_CSPAN (h) 1118_30_CNN (i) 1120_30_PBS



(j) 1127_30_ABC (k) 1129_30_CNNHL (l) 1201_30_CNBC

Figure 4.4: Speech time per speaker in each show of the RT-04F *dev1* and *dev2* datasets.

implementation performs better than the global BIC within both the **c-bic** and **c-sid** systems.

A similar result was observed in [Tranter and Reynolds, 2004]. This result may be due to a mismatch between the BIC model and the real distribution of the data. Thus only the local criterion is used in the remaining experiments.

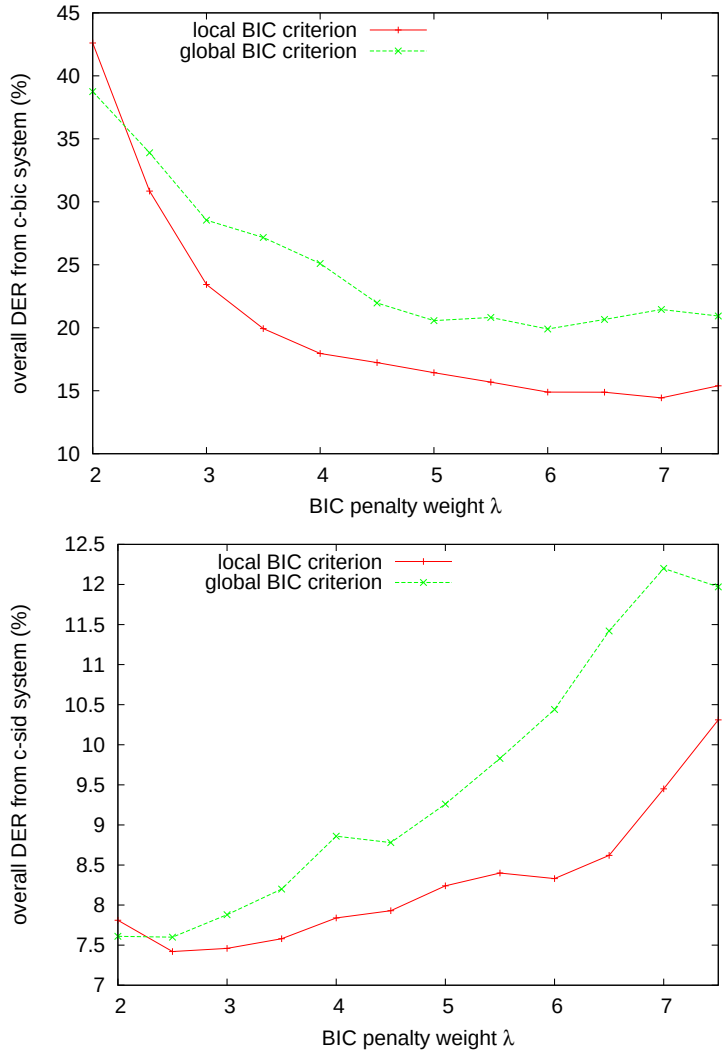


Figure 4.5: The overall diarization error on RT-04F *dev1* and *dev2*, as a function of the BIC penalty weight λ for the local and global BIC criteria. The first graph is using the **c-bic** system and the second the **c-sid** system with the SID clustering threshold δ set to 0.1.

The penalty weight λ is a critical factor to clustering performance, since it interferes in the BIC-based stopping criterion. Based on Equation 4.3 for calculating ΔBIC value, a higher penalty factor λ tends to produce more negative ΔBIC values, thus leading more segments to be merged. This means that increasing the factor λ would reduce the final cluster number. As mentioned previously, the BIC clustering stage is optimized to be under-clustered in the **c-sid** system so as

to give the SID clustering more cluster candidates to merge further, rather than being optimized directly to provide the best diarization performance in the **c-bic** system. Therefore, the optimal value of the BIC penalty weight λ for the **c-bic** system is generally greater than that for the **c-sid** system. This has been confirmed on the RT-04F development data. As shown in Figure 4.5, the minimal DER is obtained by setting the λ in the local BIC criterion to 7.0 for the **c-bic** system and 2.5 for the **c-sid** system.

4.3.5 SID clustering threshold

The effect of the SID clustering threshold δ on the speaker match error and the cluster purity error was measured on both the *dev1* and *dev2* data. A lower threshold reduces the number of final speaker clusters. As shown in Figure 4.6, reducing the threshold in the positive range results in a decrease of the match error rate with almost no degradation of the cluster purity. Moreover, there is a large range of thresholds around zero with a low speaker match error. The cluster purity error shows the speaker match error that could be achieved with the best clustering decision, as explained in section 3.1.2.

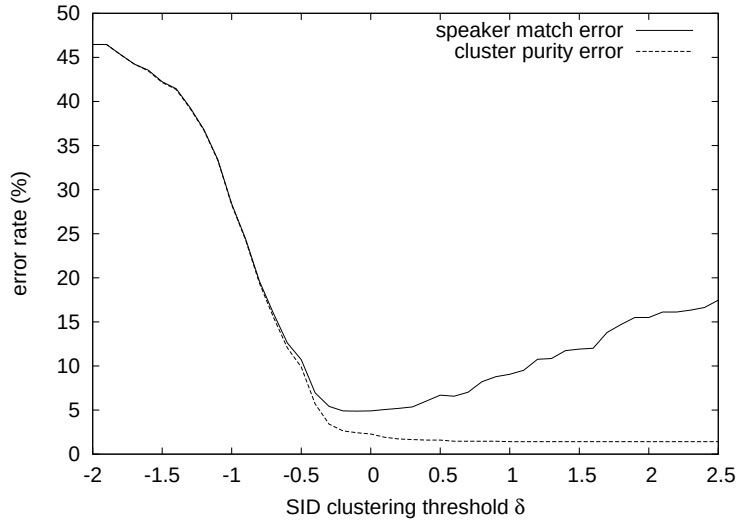


Figure 4.6: Speaker match error and purity error rates on RT-04F *dev1* and *dev2* for the **c-sid** system as a function of the SID clustering threshold δ, where the BIC penalty weight λ was set to 3.5.

4.3.6 Feature warping effects

The effects of the feature warping technique were investigated on both RT-04F *dev1* and *dev2* datasets. Since the feature warping was only performed in the SID clustering stage in the standard **c-sid** system (called “fw_sid”), the contrast experiments were designed to carry out the feature

warping procedure on both the BIC and SID stages (referred to as "fw_(bic+sid)") and not to run it at all in both cases (referred to as "nofw"). Table 4.4 shows the results of running feature warping in different clustering stages.

The standard "fw_sid" system provides 7.4% overall DER with the clustering parameters set as $\lambda = 2.5, \delta = 0.1$. Performing the feature warping on both BIC and SID stages with the same parameter settings gives a much higher cluster purity error of 6.6%, with the overall DER increased to 10.5%. A relative DER reduction of about 30% is obtained for the "fw_(bic+sid)" system by optimizing the λ to 1.5. This is a similar result to that was obtained from the "fw_sid" system, with a slight 0.2% reduction on DER.

Without feature warping in both BIC and SID stages, the diarization performance is degraded considerably with a very poor cluster purity of 14.7%, resulting in 19% overall DER. With the same parameter setting (i.e. $\lambda = 2.5, \delta = 0.1$) , performing feature warping in the SID clustering stage provides a relative DER reduction up to 61%, Although adjusting δ to 1.4 reduces the diarization error rate to 12.5% for the "nofw" system, there is still a performance margin between the "fw_sid"and "nofw" systems. This experiment demonstrates that the feature warping technique is very crucial to the performance of clustering stage.

system	*purity error (%)*	*coverage error (%)*	*speaker match error (%)*	*overall DER (%)*
fw_sid ($\lambda = 2.5, \delta = 0.1$)	1.2	4.4	4.9	7.4
fw_(bic+sid) ($\lambda = 2.5, \delta = 0.1$)	6.6	3.0	7.9	10.5
fw_(bic+sid) ($\lambda = 1.5, \delta = 0.1$)	2.1	3.7	4.7	7.2
nofw ($\lambda = 2.5, \delta = 0.1$)	14.7	3.2	16.5	19.0
nofw ($\lambda = 2.5, \delta = 1.4$)	4.5	6.3	10.0	12.5

Table 4.4: Diarization error rates of performing feature warping at different clustering stages on RT-04F *dev1* and *dev2* datasets.

4.3.7 Iteration count of MAP adaptation

It has been found that using an iterative MAP adaptation of a UBM model is effective for speaker recognition [Vogt *et al.*, 2003]. This technique updates the Gaussian posteriori probabilities at each iteration of the MAP adaptation using the data reassigned to the Gaussian components estimated from the previous iteration. The effects of iteration count of MAP adaptation were studied on the RT-04F *dev1* dataset using the **c-sid** system with the BIC penalty weight λ set to 3.5. Table 4.5 shows diarization results while changing the number of iterations for MAP adaptation from 1 to 7, with the corresponding SID clustering threshold δ selected a posterior to give the best performance. This theoretically best clustering performance is obtained by performing all possible merges without any stopping criterion and calculating the diarization error after each clustering iteration. It can be observed that further iterations provide no significant improvements in performance (only a speaker match error change of 0.7% between 1 and 7 iterations),

while the SID stage is performed much faster with 1 iteration. Furthermore, the optimal value of the SID threshold δ diminishes as the iteration number of the MAP adaptation increases. The number of iterations of MAP adaptation was set to 5 in the RT-04F evaluation.

iteration nb	*purity error (%)*	*coverage error (%)*	*speaker match error (%)*	*threshold* δ
1	2.0	3.6	4.8	0.31
2	2.4	3.2	4.7	-0.06
3	2.6	2.7	4.4	-0.21
4	2.6	2.6	4.2	-0.30
5	2.6	2.6	4.2	-0.38
6	2.6	2.6	4.1	-0.39
7	2.6	2.6	4.1	-0.42

Table 4.5: Results of different iteration counts for the MAP adaptation in the SID clustering stage on RT-04F *dev1* dataset.

4.3.8 RT-04F evaluation results

The trends observed on the development data were confirmed on the RT-04F evaluation data. As seen in Table 4.6, the results submitted to the evaluation show a slight increase in overall diarization error to 17% for the **c-bic** system and to 9.1% for the **c-sid** system. The final SAD post-processing stage gives an improvement of 0.6%, mainly by reducing false alarms in speech detection. As mentioned in [Reynolds and Torres-Carrasquillo, 2005], LIMSI's primary diarization system (i.e. **p-asr** system) had the best performance of all the participants in the RT-04F evaluation by a significant margin.

system	*missed speaker error (%)*	*false alarm speaker error (%)*	*speaker match error (%)*	*overall DER (%)*
results submitted to RT-04F				
c-bic ($\lambda = 5.5$)	0.4	1.8	14.8	17.0
c-sid ($\lambda = 3.5, \delta = 0.1$)	0.4	1.8	6.9	9.1
p-asr	**0.6**	**1.1**	**6.8**	**8.5**
post-evaluation results				
c-bic ($\lambda = 7.0$)	0.4	1.8	16.6	18.8
c-sid ($\lambda = 2.5, \delta = 0.1$)	0.4	1.8	7.0	9.2
c-sid ($\lambda = 3.5, \delta = 0.4$)	0.4	1.8	6.0	8.2

Table 4.6: Performances of **c-bic**, **c-sid** and **p-asr** systems on the RT-04F evaluation data.

Some results of the post-evaluation experiments are also given in Table 4.6. The optimal values of the BIC penalty weight λ on the RT-04F development data reported in Section 4.3.4 (i.e.

7.0 for the **c-bic** system and 2.5 for the **c-sid**) are also examined on the test data. However, these optimal values did not provide better performances compared to those using the default parameter settings. In details, the **c-bic** system with the optimal λ of 7.0 has an overall DER of 18.8%, a relative increase of about 11% compared to that obtained by setting λ to 5.5; for the **c-sid** system, the optimal and default values of λ provide similar results, with only 0.1% DER difference absolutely. These facts indicate that the optimal value of the BIC penalty weight λ is different for the development and test datasets and need to be tuned respectively. The post-evaluation results show also the speaker error for the **c-sid** system could be reduced from 6.9% to 6.0% using an SID clustering threshold δ set to 0.4, with the default λ value of 3.5. This demonstrates that the optimal SID clustering threshold varies over the different datasets and is closely related with data characteristics. Generally speaking, the parameter tuning is very critical for the performance of the diarization systems and depends largely on the show nature.

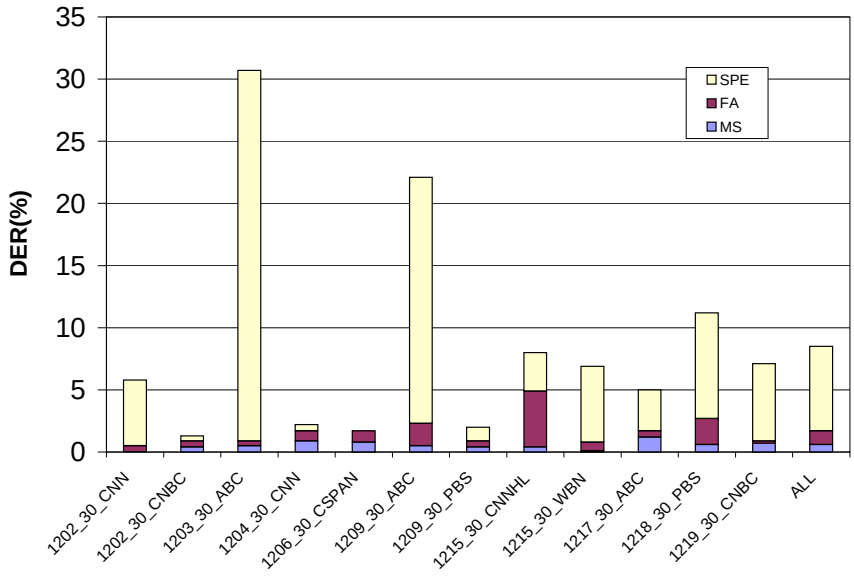


Figure 4.7: Per-show and total diarization error rates from **p-asr** system on the RT-04F evaluation dataset.

The per-show results for the **p-asr** system with the lowest DER of the RT-04F evaluation are given in Figure 4.7. Like the results obtained on the development data, there is a considerable variation in performance over the shows, from the lowest DER of 1.3% for the show "1202_30_CNBC" to the highest DER of 30.7% for the show "1203_30_ABC". Most of this variability comes from the speaker match error component due to over or under-clustering data, while the speech detection error is relatively stable. The main implication is that the parameter settings of diarization systems need to be adjusted to the different types of show which differ in the number of speakers, the dominance of speakers, the speaker turn duration and the style of the show, etc. It should be possible to improve the performance of diarization systems by automatically detecting the show characteristics and modifying the parameters accordingly.

4.4 ESTER experiments

In 2005, the multi-stage speaker diarization discussed in Section 4.2 was also used to participate in the French ESTER Broadcast News evaluation [Galliano *et al.*, 2005], where this system provided the best speaker diarization performance of the all submitted systems. The experiments conducted on the ESTER databases from the multi-stage system will be presented in this section. The default system configurations are the same settings as that were used in the RT-04F evaluation: i.e. $\lambda = 5.5$ for the **c-bic** system and $\lambda = 3.5, \delta = 0.1$ for the **c-sid**.

4.4.1 Database description

The data used in ESTER were extracted from French radio broadcast news shows, provided by ELDA (Evaluations and Language resources Distribution Agency) and the DGA (Délégation Générale pour l'Armement). The training corpus contains totally 82 hours of data from the France Inter, France Info, RFI and RTM radio stations, recorded in 1998, 2000 and 2003, with the audio file durations ranging from 10 minutes to 1 hour. The development corpus contains 8 hours of data, in 14 audio files recorded from April to July 2003 from the same stations as the training corpus. The evaluation corpus is comprised of 18 audio files recorded from October to December 2004, with a total duration of 10 hours. The evaluation corpus contains data from two radio stations ('France Culture' and 'Radio Classique') not present in the training or development corpora. There is also a 14 month interval between the recording period of the development and test data (July 2003 to October 2004) of these two corpora. Both data sets present a large variability in audio file durations (10 minutes to 1 hour). A summary of all databases used in the ESTER evaluation is given in Table 4.7.

source	*train*	*development*		*evaluation*	
	duration	*duration*	*epoch*	*duration*	*epoch*
France Inter	33 hr.	2×1 hr.	2003	2×1 hr.	2004
France Info	8 hr.	2×1 hr.	2003	1×1 hr.+2×30 min.	2004
RFI	23 hr.	2×1 hr.	2003	4×30 min.	2004
RTM	18 hr.	8×(10-20 min.)	2003	7×(14-22 min.)	2004
France Culture	-	-		1×1 hr.	2004
Radio Classique	-	-		1×1 hr.	2004

Table 4.7: The databases used in the ESTER evaluation.

4.4.2 Results on ESTER development data

For the French ESTER data, SAD was performed using the same American English speech/non-speech acoustic models as were used for the RT-04F evaluation, plus an additional speech over

music model trained on French broadcast news data for a better recognition of the jingles found in the French radio data, and the UBMs employed in the SID clustering stage were trained on the same English Broadcast News data as that were used for the RT-04F evaluation. The optimal threshold for the SID clustering on development data was $\delta = 1.5$. As can be seen in Table 4.8, the **c-sid** system has low purity and coverage error rates (4.7% and 5.2% respectively) on the ESTER development data. A 50% reduction of the overall error rate is gained by adding the **c-sid** system to the **c-bic** system. The δ threshold found for the RT-04F data (i.e $\delta = 0.1$) did not carry over to the ESTER data. This may be due to the larger variability in show sources, durations and types observed in ESTER and the mismatch between the UBMs and the ESTER data.

system	*purity error (%)*	*coverage error (%)*	*overall DER (%)*
c-bic ($\lambda = 5.5$)	7.2	10.6	15.8
c-sid ($\lambda = 3.5, \delta = 1.5$)	4.7	5.2	8.0

Table 4.8: The purity, coverage and overall diarization error rates from the **c-std**, **c-bic** and **c-sid** systems on the ESTER development dataset.

4.4.3 SID clustering threshold

Figure 4.8 demonstrates the speaker match error and the cluster purity error as a function of the SID clustering threshold for the **c-sid** system. This is a similar result as that was observed on the RT-04F development data (c.f. Figure 4.6), with the BIC penalty weight set to 3.5 in both cases: the speaker match error curve decreases with a stably low cluster purity error when reducing the threshold δ in a positive range, and then it increases fast with the same tendence for purity error when continuing reducing the δ. However, the minimum speaker match error is obtained with different δ value on two development datasets (i.e. around zero for the RT-04F and around 1.5 for the ESTER). Therefore, the best threshold for stopping the SID clustering depends largely on the specific data. In addition, it can be found that the minimum speaker match error on the RT-04F data is less than that can be achieved on the ESTER data. The possible reason is that the identical UBMs trained on the English broadcast news data for RT-04F were directly used on the ESTER development data.

4.4.4 ESTER evaluation results

Results on the ESTER evaluation data are given in Table 4.9, with the parameter setting optimized on the development data. The overall diarization error was reduced from 13.8% for the **c-bic** system to 11.5% for the **c-sid** system. The submitted system also had the best performance for the speaker diarization task in the ESTER evaluation [Galliano *et al.*, 2005]. It needs to note that these scorings are computed over all the transcribed audio except long regions of music and publicity. As long silence segments are also evaluated in this case, the diarization system is

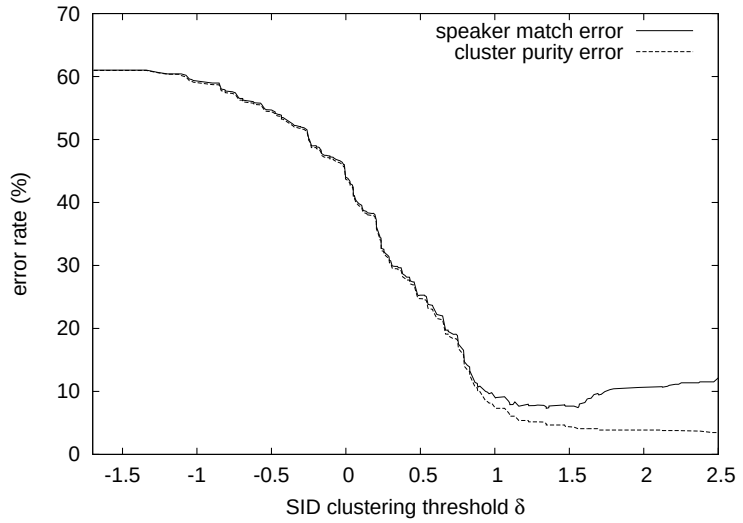


Figure 4.8: Speaker match error and purity error rates on ESTER development dataset for the **c-sid** system as a function of the SID clustering threshold δ.

required to be able to detect them, otherwise high false alarm error will be produced. However, when the performance metrics are measured strictly on labeled speaker segments and the long silence regions are not taken into account, the overall DER should be reduced, especially false alarm error (e.g. for the **c-sid** system, the DER and false alarm error are reduced to 10.2% and 0 respectively). In a post-evaluation experiment, a 20% relative reduction of the overall diarization error was observed for the **c-sid** system with the best a posteriori threshold, showing that the error rate is highly dependent on the clustering threshold.

system	*missed speaker error (%)*	*false alarm speaker error (%)*	*speaker match error (%)*	*overall DER (%)*
results submitted to ESTER				
c-bic ($\lambda = 5.5$)	0.7	1.0	12.1	13.8
c-sid ($\lambda = 3.5, \delta = 1.5$)	0.7	1.0	9.8	11.5
post-evaluation results				
c-sid ($\lambda = 3.5, \delta = 2.0$)*	0.7	1.0	7.4	9.1

Table 4.9: Performances of **c-bic** and **c-sid** systems on the ESTER evaluation data.

For the submitted **c-sid** system with the best performance, the per-show and total diarization errors on the ESTER evaluation data are given in Figure 4.9. Like the RT-04F evaluation results (see Figure 4.7), there is also a large variability of the system performance over the shows with different durations. To investigate whether there are some correlations between the acoustic characteristics of the data and the diarization error rates, the system output for the show

"1026_30_RFI" with the highest speaker match error of approximately 30% was analyzed. It has been found that this 30-minute show contains a 20-minute spot coverage to an American evangelic community on the American presidential campaign, where the same reporter occurs in the presence of different background noise, music or noisy background speech from different persons and the English speech from the interviewees is overlapped totally or alternately by French translation. Due to this complexity in the acoustic environments, the diarization system tends to divide speech from speakers into several clusters corresponding to different acoustic environments and provide a higher speaker match error.

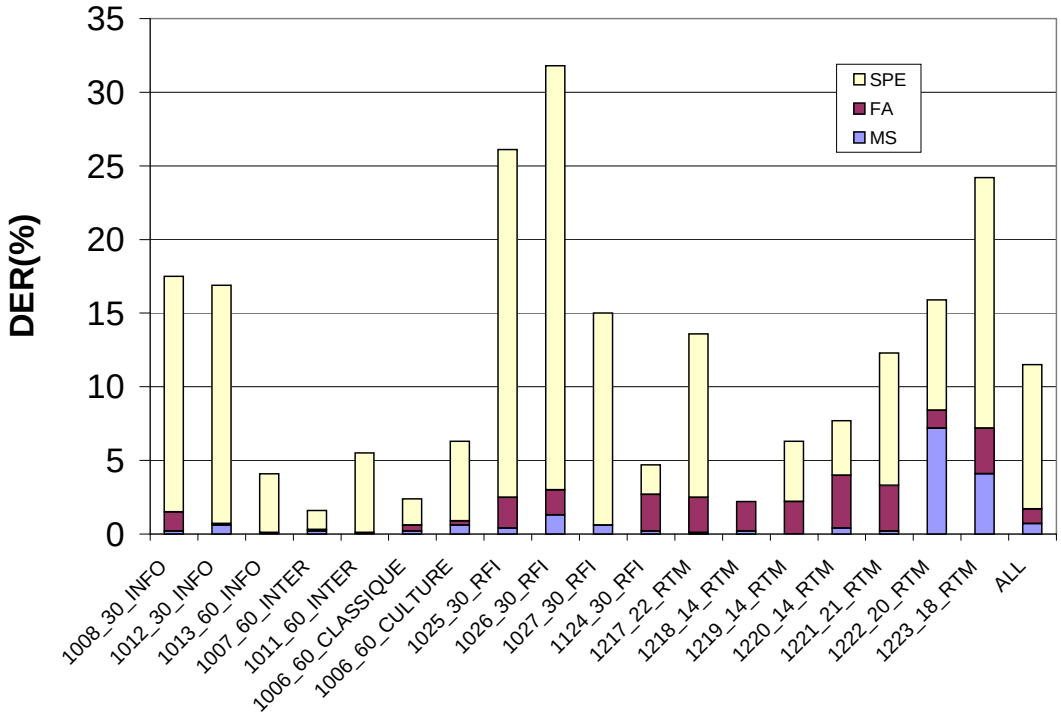


Figure 4.9: Per-show and total ESTER evaluation results from the **c-sid** system.

4.5 Robustness of the multi-stage diarization system

The experiments conducted on both RT-04F and ESTER data show that diarization performance depends largely on the setting of the BIC penalty weight λ and the SID clustering threshold δ. In order to improve the robustness of the multi-stage diarization system, the optimal values of the system parameters (i.e. λ and δ) are studied on the shows with different durations. To do this, a subset of 20 1-hour BN shows (denoted as *'1hour'*) extracted from the ESTER training data serves as the examined data, all these shows are emitted by the radio station "France Inter" and collected during two complete weeks (i.e. 10 weekdays), with two 1-hour shows on each day. In

addition, two other subsets are also tested for comparing diarization performance: *'first30min'* consisting of the first 30 minutes BN shows and *'second30min'* composed of the second half of the shows.

4.5.1 BIC penalty weight λ vs. show duration

In the presented multi-stage diarization system (c.f. Section 4.2), the BIC clustering stage is configured to provide an under-clustered input to the second SID clustering. To investigate the correlation between the value of the BIC penalty weight λ and the duration of BN shows, the BIC clustering stage is performed without any stopping threshold, i.e. repeating the process of cluster merging until only one cluster is left. Therefore, the best BIC clustering performance can be achieved by setting the stopping threshold a posteriori, i.e. the optimal threshold is equal to the ΔBIC value of the two clusters whose merging generates a clustering solution with the minimum speaker match error.

λ	*1hour*		*first30min*		*second30min*	
	SPE (%)	*threshold*	*SPE (%)*	*threshold*	*SPE (%)*	*threshold*
1.0	12.30	3037	9.63	2592	5.78	2767
1.5	11.31	2710	9.79	2349	5.77	2470
2.0	10.34	2330	9.48	2151	5.42	2427
2.5	8.49	2075	9.22	1843	5.12	1446
3.0	8.46	1610	9.07	1091	4.34	1030
3.5	8.38	841	9.43	810	4.27	728
4.0	9.84	592	9.77	560	4.44	7
4.5	13.47	342	11.33	315	5.37	-580
5.0	17.59	-238	12.41	-256	6.21	-831
5.5	20.92	-529	14.01	-367	6.62	-1052
6.0	22.97	-990	15.60	-694	7.88	-1427

Table 4.10: Speaker match error (SPE) obtained on the datasets *1hour*, *first30min* and *second30min* using different BIC penalty weight λ values, where the threshold is selected a posteriori as the ΔBIC value that corresponds to the best clustering solution.

The BIC clustering results using different values of the penalty weight λ on the datasets *1hour*, *first30min* and *second30min* are given in Table 4.10, where the speaker match errors are obtained using an a posteriori optimal threshold to stop the BIC clustering process. For the datasets *1hour* and *second30min*, setting the BIC penalty weight λ to 3.5 provides the best diarization results (8.38% for *1hour* and 4.27% for *second30min*). For the *first30min* dataset, the lowest speaker match error of 9.07% is observed by setting λ to 3.0. However, in the three cases, the corresponding optimal thresholds for stopping the BIC clustering are not zero as is adopted in the multi-stage diarization system. The speaker match error depends largely on the value of penalty weight λ,

but the optimal value of λ seems to not relate closely with the duration of audio recording.

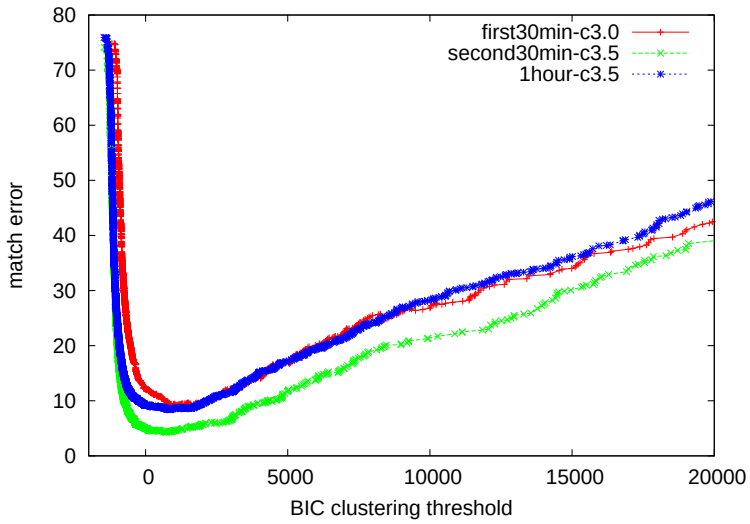


Figure 4.10: The speaker match error obtained using the optimal penalty weight λ as a function of the BIC clustering threshold on the datasets *1hour*, *first30min* and *second30min*.

To further examine the trend of clustering performance during the cluster merging process, Figure 4.10 demonstrates the speaker match error as a function of the BIC clustering threshold, where the BIC penalty weight λ is set to 3.5 for both *1hour* and *second30min* datasets and 3.0 for the *first30min* dataset. The speaker match error curve shows the similar tendence on the three datasets: the speaker match error reduces rapidly with the increase of the the BIC value in the negative range and continues diminishing in a small positive scope close to zero, afterward it begins to increase monotonously. This phenomenon conforms to the theory of the BIC criterion that a negative value of ΔBIC indicates that two sets of data are better drawn from two distinct models than a single model. Therefore, the theoretical stopping threshold of the BIC clustering would be zero. As shown in Figure 4.10, the ΔBIC value associated with the lowest match error on the dataset *second30min* is closer to 0 than those for *1hour* and *first30min*, thus better clustering performance is obtained on the *second30min* dataset.

The optimal BIC clustering threshold as a function of the penalty weight λ is shown in Figure 4.11. Similar results are obtained on three datasets *1hour*, *first30min* and *second30min*. The optimal value of the BIC clustering threshold reduces with the increase of the penalty weight λ and seems to linearly correlate with the λ value, approximately formulated as $3800 - 800\lambda$ based on the results given in Figure 4.11. With the consideration of this correlation, the BIC clustering stopping criterion can be generally expressed as:

$$-\log R - \lambda P > a\lambda + b \tag{4.7}$$

The above equation can be further reformulated as :

$$-\log R - \lambda(P + a) > b \tag{4.8}$$

where R is likelihood ratio of the test that two compared data are uttered by the same speaker or different speakers (c.f. Equation 2.28) and the penalty P is computed according to the Equation 4.4. In the initial local BIC criterion, the penalty term relates only to the dimensionality of acoustic vectors and the size of two clusters to be merged. However, the results given in Figure 4.11 imply that the initial BIC criterion needs some modifications in the penalty modeling, e.g. introducing an additional factor a in the penalty term. Although this additional penalty factor is experimentally found not to correlate directly with the duration of audio recordings, the experiment presented here inspires the work in the future to improve the BIC method by studying the elements that may influence the penalty factor.

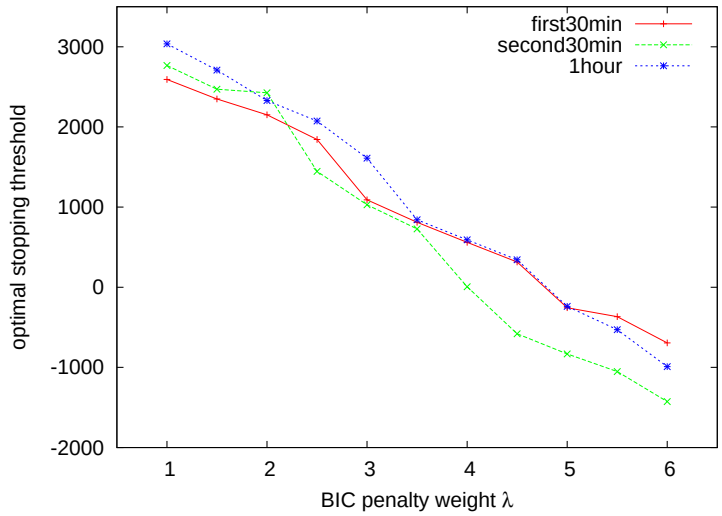


Figure 4.11: Optimal BIC clustering threshold on the datasets *1hour*, *first30min* and *second30min* as a function of the penalty weight λ.

4.5.2 SID clustering threshold δ vs. show duration

The SID clustering without the use of a stopping threshold is also performed on the datasets *1hour*, *first30min* and *second30min*. The Figure 4.12 demonstrates the speaker match error obtained on the three datasets as a function of the SID clustering threshold, with the BIC penalty weight λ set to 3.5 for all datasets. The speaker match error curve shows a similar tendency on the datasets consisting of the BN shows with different durations. For the two datasets that are composed of the 30 minutes shows, a relatively large difference is found between the optimal values of the threshold δ, while the optimal δ values for the dataset *1hour* and *first30min* are

closer to each other. However, in general, smaller optimal value of the threshold δ gives the best clustering performance on the shows with shorter duration. On the three datasets, lowest speaker match error is obtained when the value of δ ranges between 1.0 and 1.5. The similar diarization performance is achieved on the ESTER development data with δ set to 1.5 (c.f. Table 4.8).

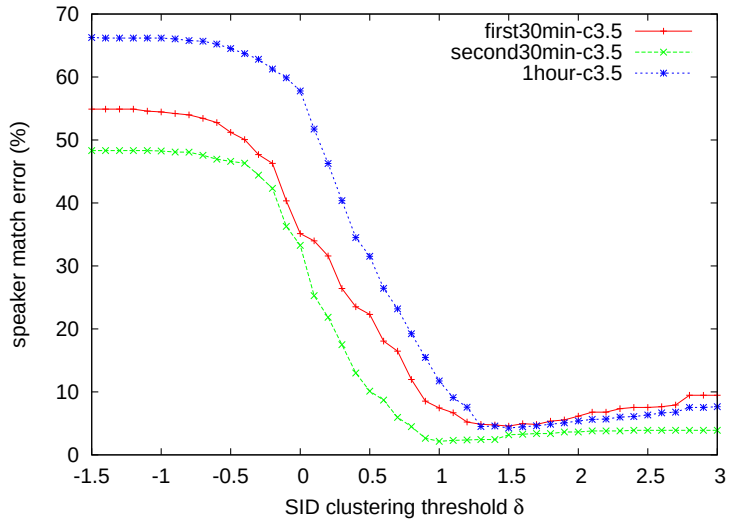


Figure 4.12: Speaker match error on the datasets *1hour*, *first30min* and *second30min* as a function of the SID clustering threshold δ.

4.5.3 Conclusions on the robustness experiments

The experiments described in this section investigate the correlation between the parameter setting of the multi-stage diarization system (i.e. BIC penalty weight λ and SID clustering threshold) and the duration of BN shows. The results obtained on the datasets *1hour*, *first30min* and *second30min* indicate that the duration of audio recordings is not a crucial factor affecting the parameter tuning. However, the experiments carried out with the BIC clustering stage show that the penalty term in the BIC criterion may be revised by incorporating the factors that reflect some other characteristics of audio recordings. This is a preliminary study on the robustness of the diarization system, many further investigations remain to complete, such as examining different dimensions of MFCCs etc.

4.6 Conclusions

The baseline and improved multi-stage speaker diarization systems for Broadcast News have been described in this chapter. The improved system builds upon a baseline speaker partitioning

system which had been optimized for the automatic transcription task. Considering the constraints of the speaker diarization task are different, several modifications to the baseline system have thus been explored. First, the iterative GMM clustering was replaced by an agglomerative BIC clustering, using single full-covariance Gaussian models. A local BIC merging and stop criterion was shown to outperform the global criterion. A second clustering module was then applied to the output of the BIC clustering system, relying on techniques used for speaker identification and verification: acoustic channel normalization, and MAP adaptation of a reference GMM with a large number of Gaussians.

The improved multi-stage system was demonstrated to perform much better than the baseline system for the diarization task. On the RT-04F development data, a relative error reduction of over 70% was achieved when compared to the baseline system. This multi-stage system produced the best diarization performance in both the NIST RT-04F and the ESTER evaluations by a significant margin. An overall diarization error rate under 10% was obtained on the RT-04F evaluation data, while the DER from the single-stage BIC clustering system was over 15%. Focusing on the speaker match error only, the multi-stage system provides an reduction of up to 50% relative to the BIC clustering system. Similar results were also obtained on the ESTER evaluation data, with a relative speaker error reduction of approximately 40%. These dramatic improvements over the baseline system result from several changes: the combination of two different clustering stages, each one focusing on a different acoustic aspect with more complex modeling in the second stage, and the use of acoustic channel normalization methods suited to speaker identification. A system following this architecture developed at Cambridge University demonstrated similar improvements [Sinha *et al.*, 2005], where it was observed that a very important part of the gain was obtained by the feature warping normalization.

Chapter 5

From Broadcast News to meetings

The adaptations of the LIMSI broadcast news speaker diarization system to meetings are presented in this chapter. The baseline system provided an overall diarization error of 8.5% on broadcast news data in the NIST Rich Transcription 2004 Fall (RT-04F) evaluation. However, since it produces a high missed speech error rate on lecture data, a different Speech Activity Detection (SAD) approach relying on the smoothed Log-Likelihood Ratio (LLR) between the speech and non-speech models was integrated into the diarization system for lectures, within the framework of the NIST Rich Transcription 2006 Spring (RT-06S) meeting recognition evaluation. This modified system gave an overall diarization error of 21.5% on the RT-06S Multiple Distant Microphone (MDM) lecture data. The improvements to the diarization system for general meetings data (i.e conferences and lectures) made during the NIST RT-07S evaluation are introduced in this chapter as well. A variance normalization technique was performed on the acoustic features within the LLR-based SAD module. A new selection of training data with the corresponding forced alignment transcription references were used to train the speech and non-speech models and UBM used in the SID clustering stage. In addition, the diarization system employed the beamformed signals which were generated by the ICSI delay&sum signal enhancement system from all available MDM channels. The LIMSI RT-07S diarization system gives comparable results on both the RT-07S conference and lecture evaluation data for MDM condition (26.1% diarization error on the conference test data and 25.8% error rate on the lecture test data).

5.1 Comparison between BN and meetings

Most research efforts on speaker diarization have been focused on the broadcast news domain, thus many methods have been proposed and their implementations have been proven to be successful in this area [Tranter *et al.*, 2004; Barras *et al.*, 2006]. In recent years, the focus of research interests has evolved towards the meeting environment [Istrate *et al.*, 2005a; Anguera *et al.*, 2006c; Wooters and Huijbregts, 2007]. NIST has been organizing the Rich Transcription evaluations for meetings in collaboration with the Augmented Multi-party Interaction (AMI)

and Computers In the Human Interaction Loop (CHIL) projects since 2005, where two different meeting sub-domains have been proposed: conference room meetings and lecture room meetings. The two sub-domains differ in the amount of the participant interactivity and microphone configurations.

As the multi-stage speaker diarization system presented in Chapter 4 has provided a state-of-the-art diarization performance on broadcast news data, it is feasible to keep the roughly same system structure and to make some necessary adaptations aiming to the new requirements of the meeting task. To do this, an analysis of the differences from broadcast news to meetings is first performed by examining the input audio conditions and the reference speaker segmentations. This analysis has been carried out on some datasets described as below:

- **RT-04F broadcast news development dataset**: as presented in Section 4.3.1, it is composed of two portions, *dev1* and *dev2*, each consisting of 6 30-minute show excerpts, extracted from multiple US radio or TV Broadcast News programs. The *dev1* is the same dataset that was used as the RT-03 evaluation data, with the shows recorded in February 2001, whilst *dev2* is a new development database for the RT-04F evaluation, whose shows were recorded in Nov/Dec 2003. The ensemble of both *dev1* and *dev2* (called as *BN dev04f*) has been used for this analysis.

- **RT-06S meeting development dataset**: there are two subsets in this dataset, the conference room meeting subset and the lecture room meeting subset. The RT-06S conference development database (denoted as *conf dev06s*) consists of 10 meeting excerpts, with each lasting about 12 minutes, recorded at 5 different sites: AMI, CMU (Carnegie Mellon University), ICSI (International Computes Science Institute), NIST and VT (Virginia Tech). The RT-06S lecture development dataset (denoted as *lect dev06s*) has a total of 35 seminar excerpts, with the duration ranging from 69 seconds to 937 seconds in each. This database is composed of all the seminars which served as the RT-05S lecture test data and several additional interactive seminars, collected at 5 of the CHIL partner sites respectively: AIT (Athens Information Technology), IBM, ITC (Istituto Trentino di cultura), UKA (University of Karlsruhe) and UPC (Universitat Politecnica de Catalunya).

For the RT-06S lecture development dataset, because the available segmentation references cover only 31 seminar excerpts, in which one seminar (c.f. Section 5.3.6) has a large amount of missed speech in the corresponding transcription, a subset consisting of 30 excerpts has thus been used for the analysis.

5.1.1 Audio input conditions

The purpose of the analysis of the audio input conditions is to show the different complexity of the input signals between broadcast news and meetings domains. This is particularly necessary for meeting data due to the use of the different types of microphone. In addition, the Speech to

Noise Ratio (SNR) will be estimated on the signals of each dataset. Although this is a simple SNR approximation with bimodal modeling of log energy, it provides coherent SNR values to allow the comparison of signal qualities between different data domains.

Regarding the broadcast news shows first, there is a single audio input for each BN show. In order to further investigate the signal quality, the estimated SNR of each show from the RT-04F development database is given in Figure 5.1. A large variation in SNR values is observed on broadcast news data, with an average SNR of 28.8 dB. This phenomenon can be explained by analyzing the organization of BN programs. On the one hand, there is normally an anchor in each broadcast news show who takes charge of the whole program, the corresponding part of the signal is recorded via a very good quality microphone in a closed studio where the acoustic environment is very well controlled. On the other hand, many shows contain some reports from the spot, in the presence of different background noise. Hence, the signal quality of broadcast news shows varies largely according to the amount of live reports.

show	*SNR*
20010206_1830_1900_ABC_WNT	70.2
20010217_1000_1030_VOA_ENG	28.2
20010220_2000_2100_PRI_TWD	22.5
20010221_1830_1900_NBC_NNW	69.5
20010225_0900_0930_CNN_HDL	71.5
20010228_2100_2200_MNB_NBW	43.7
20031115_180413_CSPAN_ENG	23.1
20031118_050200_CNN_ENG	48.7
20031120_003511_PBS_ENG	19.7
20031127_183655_ABC_ENG	22.0
20031129_000712_CNNHL_ENG	23.5
20031201_203000_CNBC_ENG	22.9
average	28.8

Table 5.1: SNR estimations on the RT-04F broadcast news development dataset.

In the framework of NIST meeting recognition evaluations, the meeting domain has been divided into two sub-domains since 2005, with different microphone setups in each sub-domain:

- **Conference room meetings**: these are carried out by several participants sitting around a meeting table, with each one wearing a high quality close-talking microphone. Since the oriented goal is meetings, all participants are engaged in the conversation with a similar interaction level.
- **Lecture room meetings**: these are also called seminar-like meetings, where a lecturer stands normally in front of all the other participants. These seminars typically consist of a presentation from the lecturer followed by a question/answering session or discussion

period. Thus the lecture meetings have less interactivity between participants compared to the conference meetings.

There are different types of microphone used in each meeting room, thus multiple audio recordings are available to provide synchronized signals for each meeting. In NIST meeting recognition evaluations, several audio input conditions were proposed for different evaluation tasks and data type. According to the RT-06S meeting recognition evaluation [NIST, 2006], the list below explains each audio input condition and indicates which sub-domain it is applied to:

- **Single Distant Microphone (SDM)**: this condition is defined as the audio recorded from a single omni directional microphone, centrally located on the meeting table. This audio recording is always included in the set of MDM channels. This condition is available for both conference and lecture meetings.
- **Multiple Distant Microphone (MDM)**: this evaluation condition contains a set of audio collected from several omni directional microphones placed on a table between the participants. It is required as the primary task for conference and lecture sub-domains.
- **Multiple Mark III Microphone Arrays (MM3A)**: in some lecture meeting rooms, the Mark III array is used to provide 64 signal channels. A beamformed signal generated by the University of Karlsruhe is available to each participating team for RT-06S data.
- **Multiple Source Localization Microphone Arrays (MSLA)**: this is a 4 digit microphone array arranged in an "T" topology. As this microphone array was built and used by the project CHIL, this condition exists only in the lecture meeting sub-domain.
- **All Distant Microphone (ADM)**: this evaluation condition permits the use of all distant microphone presented previously.
- **Individual Head Microphone (IHM)**: this condition includes the audio recordings collected from a close-talking microphone worn by some of the participants in the meeting. It was not evaluated for speaker diarization task, but used for speech-to-text and speech activity detection tasks.

Since the MDM condition is the primary evaluation condition for both conference and lecture meetings, the analysis of the audio input conditions in meeting domain has been performed on the MDM condition. The channel numbers available for MDM task and the SNR estimation are given in Table 5.2 and 5.3 for RT-06S conference and lecture meetings development datasets respectively. Compared to only a single audio recording available for a broadcast news show, the MDM condition provides at least 2 audio input for a same meeting. The MDM channels are recorded by using different distant microphones located at various positions on a meeting table, which leads to difference in arrival time from a same sound source between each of the multiple distant channels. How to effectively exploit the multiple signals from different microphones is a crucial problem for the MDM speaker diarization task. In the field of the speaker diarization

for meetings, various techniques have been proposed to use the information from all available channels [Jin *et al.*, 2004; Anguera *et al.*, 2005a; Anguera *et al.*, 2005b; Istrate *et al.*, 2005a; Pardo *et al.*, 2006a; Pardo *et al.*, 2006b].

meeting	*channels number*	*min. SNR*	*max. SNR*	*ave. SNR*
AMI_20041210-1052	12	10.6	14.0	12.6
AMI_20050204-1206	16	10.6	12.8	11.5
CMU_20050228-1615	3	9.6	10.3	10.0
CMU_20050301-1415	3	12.3	13.3	12.7
ICSI_20010531-1030	6	6.9	10.1	8.5
ICSI_20011113-1100	6	8.3	10.1	9.2
NIST_20050412-1303	7	13.7	24.4	19.4
NIST_20050427-0939	7	8.3	14.9	11.3
VT_20050304-1300	2	17.2	19.7	18.4
VT_20050318-1430	2	15.6	18.5	17.1
all	-	6.9	24.4	12.4

Table 5.2: SNR estimations on the RT-06S conference development dataset.

When looking into the details of SNR measurements, it can be found that the meeting data have lower SNR estimations than the broadcast news data. In details, the average SNR of both the conference and lecture meetings recordings are around of 12.5 dB, while the broadcast news shows have an averaged SNR value close to 30 dB. This fact demonstrates that the meeting recordings usually have less signal quality than the broadcast news ones. The possible reason is that the distant microphones employed in meeting domain commonly have lower quality than that used in the broadcast studio. Moreover, the SNR estimation varies largely across different microphone channels for some meetings (e.g. the "ITC_20050429_Segment1" seminar).

5.1.2 Speaker turn duration analysis

In addition to investigate the audio input conditions, some other studies are made on the reference transcriptions. The force aligned diarization references were used for the broadcast news and conference meetings datasets, while a set of hand-made references were used for the lecture meetings as no forced alignment transcriptions is available for this lecture database. The aligned transcriptions were derived from the manual transcriptions via the LIMSI forced alignment engine that aligns both the words and the non-lexical elements (e.g. cough, breath, lipsmack and laugh) to signals. Based on these forced alignment times, segment boundaries are identified when either a new speaker starts speaking, or a break of 0.3 seconds or more occurs during a speaker turn.

An ideal speaker segmentation algorithm is required to provide sufficiently long segments in

meeting	*channels number*	*min. SNR*	*max. SNR*	*ave. SNR*
AIT_20050726_Segment1	4	16.8	20.2	18.4
IBM_20050824_Segment1	3	7.1	7.9	7.4
ITC_20050429_Segment1	5	7.8	28.8	12.3
UKA_20041123_A_Segment1	5	7.5	16.5	13.5
UKA_20041123_A_Segment2	5	7.8	13.3	11.4
UKA_20041123_B_Segment2	5	11.3	18.5	15.8
UKA_20041123_C_Segment1	5	8.9	14.4	12.7
UKA_20041123_C_Segment2	5	10.1	17.2	15.1
UKA_20041123_D_Segment1	5	7.6	13.7	11.3
UKA_20041123_D_Segment2	5	10.9	14.7	13.2
UKA_20041123_E_Segment1	5	7.8	12.2	10.3
UKA_20041123_E_Segment2	5	8.0	12.1	10.6
UKA_20041124_A_Segment1	5	7.2	16.2	12.9
UKA_20041124_B_Segment1	5	7.7	16.3	13.4
UKA_20041124_B_Segment2	5	7.8	13.0	11.5
UKA_20050112_Segment1	5	12.2	15.6	13.4
UKA_20050112_Segment2	5	9.9	13.6	12.1
UKA_20050126_Segment1	5	12.2	18.8	16.1
UKA_20050127_Segment1	5	9.2	14.3	11.4
UKA_20050128_Segment1	5	8.5	15.5	13.3
UKA_20050128_Segment2	5	11.9	19.3	17.2
UKA_20050202_Segment2	5	10.5	18.7	14.3
UKA_20050209_Segment1	5	7.6	13.1	11.3
UKA_20050209_Segment2	5	7.2	12.9	11.2
UKA_20050310_A_Segment1	5	11.0	16.9	14.6
UKA_20050310_A_Segment2	5	10.4	16.2	13.7
UKA_20050310_B_Segment1	5	10.3	16.7	14.1
UKA_20050314_Segment1	5	8.3	14.0	11.7
UKA_20050314_Segment2	5	7.1	10.4	9.4
UPC_20050601_Segment1	4	12.6	12.8	12.7
all	-	7.1	28.8	12.9

Table 5.3: SNR estimations on the RT-06S lecture development dataset.

order to construct robust models to the subsequent clustering stage, at the same time, it needs to generate pure segments, i.e. each containing only one speaker. The threshold to detection decisions is very crucial for speaker segmentations and it depends largely on the data type. For that reason, the speaker turn durations are examined on the broadcast news and meetings datasets. To compute this feature, speech segments from the same speaker that contain time intervals of

less than 2 seconds are bridged into a single speaker turn, then its corresponding duration is counted.

Table 5.4 shows the minimum, maximum and average speaker turn durations in the RT-04F broadcast news and RT-06S conference and lecture meetings development data. The average speaker duration of the conference meetings is much shorter (only 5.6 seconds) than that of the broadcast news and lectures. This reflects that a high level of interactions exists between conference participants or there are many speaker pauses of more than 2 seconds between speech segments from the same speaker. The speaker turn duration varies greatly across the different excerpts in the broadcast news and lectures domains. Moreover, the fact that one lecturer speaks mainly in a lecture leads to a maximum average speaker turn duration of 40.7 seconds in lecture room meetings. It should be noticed that the human-labeled transcription used for the lectures contains longer speaker segments than the forced alignment ones, which can also result in a greater average length of speaker turns.

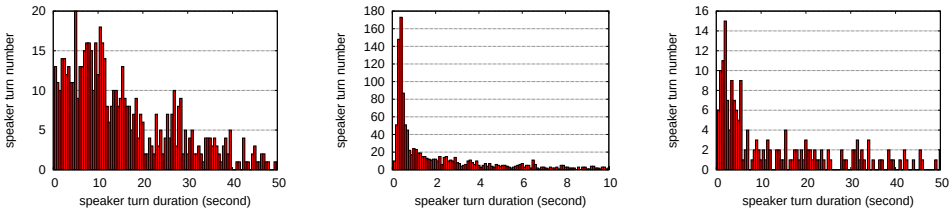
dataset	*total audio duration*	*turn number*	*min. turn duration*	*max. turn duration*	*ave. turn duration*
BN dev04f	24040.7	675	0.05	337.4	22.2
conf dev06s	22030.0	1383	0.07	95.5	5.6
lect dev06s	11364.3	250	0.24	309.2	40.7

Table 5.4: Average speaker turn duration in the broadcast news and conference and lecture meetings datasets (in second).

Figure 5.1 illustrates the distribution of the speaker turn durations for the three different types of data. For the broadcast news and lectures data, the histogram displays the speaker turns lasting less than 50 seconds, with a resolution of 0.5 seconds, while the histogram of the conference meetings presents the speaker turns of the first 10 seconds using a resolution of 0.1 second. The conference meetings contains a large number of speaker turns with the duration around 0.4 seconds. These small speaker turns are possibly derived from the overlapping speech existing in the conference data. The speaker turn distributions on both the broadcast news and lectures are relatively uniform compared to the conferences.

5.1.3 Total speech time per speaker analysis

The total speech time of each speaker in an audio recording is also investigated in both broadcast news and meetings domains. This is beneficial to regroup the segments from the same speaker, as the stopping criterion of the speaker clustering could take into account the average cluster size that needs to be adjusted according to the average speaker duration in the specific domain. The maximum, minimum and average total speech time per speaker as well as the speaker number are given in Table 5.5, 5.6 and 5.7 for RT-04F broadcast news development shows, RT-06S conference and lecture development meetings respectively.



(a) RT-04F BN dev (b) RT-06S conf dev (c) RT-06S lect dev

Figure 5.1: Speaker turn duration histograms on the broadcast news and conference and lecture meetings data.

show	*max. speaker duration*	*min. speaker duration*	*ave. speaker duration*	*speaker number*
20031115_180413_CSPAN_ENG	783.5	54.1	464.0	3
20031129_000712_CNNHL_ENG	520.5	12.5	119.3	9
20010228_2100_2200_MNB_NBW	456.9	6.6	120.5	10
20031201_203000_CNBC_ENG	303.4	0.5	88.0	11
20010225_0900_0930_CNN_HDL	617.4	2.5	72.6	16
20031118_050200_CNN_ENG	427.6	2.2	57.6	17
20010217_1000_1030_VOA_ENG	310.2	4.6	79.4	20
20010221_1830_1900_NBC_NNW	286.8	3.7	52.9	21
20031127_183655_ABC_ENG	188.5	1.0	38.3	23
20010220_2000_2100_PRI_TWD	236.0	2.4	61.0	25
20010206_1830_1900_ABC_WNT	347.1	1.1	40.9	27
20031120_003511_PBS_ENG	246.2	0.4	46.4	27
all	783.5	0.4	68.1	209

Table 5.5: Average total speech time per speaker in each show of the RT-04F BN development dataset (in second).

On the broadcast news data, although the average speaker duration over all the speakers is 68.1 seconds, it varies largely across the shows, from a minimum of 38.3 seconds to a maximum of 464.0 seconds. It is found that the shows containing more speakers usually have shorter speaker duration in average. Since each show has a similar duration of about 30 minutes, this is possible that the different speaker number in each show leads to the variability in the average speaking time. In the case of meetings, both conference and lecture room meetings have a longer average speaker duration than that of the broadcast news shows, even though there is a duration difference of approximately 20 seconds between the conference and lecture meetings (i.e. 117.8 and 134.1 seconds for the conference and lecture sub-domains respectively).

Table 5.5, 5.6 and 5.7 exhibit also the minimum speaker duration in the broadcast news, conference and lecture meetings respectively. As can be seen from these tables, most broadcast news shows and some lecture meetings contain speakers with a very short speech time (i.e. less than 5 seconds). It is common for BN shows to contain several public interviewees that just give some simple expression in certain reports. This type of speakers are difficult to model due to the lack of available data, therefore speaker diarization systems tends to regroup these speakers with other speakers who have similar vocal characteristics. Since a time-based metric is used in NIST evaluations to measure the speaker diarization performance, most diarization systems are designed to find out major speakers, instead of all speakers.

meeting	*max. speaker duration*	*min. speaker duration*	*ave. speaker duration*	*speaker number*
AMI_20041210-1052	310.2	37.2	143.1	4
AMI_20050204-1206	212.6	87.1	141.5	4
CMU_20050228-1615	230.5	133.8	174.4	4
CMU_20050301-1415	361.2	54.4	149.6	4
ICSI_20010531-1030	331.2	15.3	116.2	5
ICSI_20011113-1100	119.6	5.7	73.7	9
NIST_20050412-1303	211.9	12.7	96.4	6
NIST_20050427-0939	330.0	21.2	145.5	4
VT_20050304-1300	227.9	13.4	113.7	5
VT_20050318-1430	147.5	40.3	96.6	5
all	361.2	5.7	117.8	50

Table 5.6: Average total speech time per speaker in each RT-06S conference development meeting (in second).

The maximum speaker duration indicates the total speech time from the dominant speaker in a given audio recording. For the broadcast news data, the maximum speaker duration is much higher than the average one, while there is less difference between the maximum and average speaker duration for the lecture meeting data. This is because each BN show usually contains an anchor who coordinates the newscast, it is clear that this speaker speaks much more than the others. On lecture meeting data, a main speaker is also found in the meetings contributed by AIT, IBM, ITC and UPC and some meeting excerpts from UKA. This is derived from the fact that the most part of a lecture meeting includes the talk from a lecturer, succeeded by a discussion or question/answer section. However, this is not the case for all UKA lecture excerpts, since one UKA seminar is usually divided into several excerpts for the evaluation purpose. For instance, the first half of a lecture is used to form an excerpt, thus only one speaker existing in it, and the rest is used as another excerpt, thus consisting of several speakers. The less difference between the maximum and average speaker duration on this conference meeting set implies that a relatively higher interaction exists between conference participants.

meeting	*max. speaker duration*	*min. speaker duration*	*ave. speaker duration*	*speaker number*
AIT_20050726_Segment1	590.7	5.3	164.4	4
IBM_20050824_Segment1	722.7	13.7	368.2	2
ITC_20050429_Segment1	816.3	20.0	286.2	3
UKA_20041123_A_Segment1	280.6	280.6	280.6	1
UKA_20041123_A_Segment2	70.4	30.3	50.4	2
UKA_20041123_B_Segment2	70.5	4.0	27.9	3
UKA_20041123_C_Segment1	269.4	269.4	269.4	1
UKA_20041123_C_Segment2	187.4	4.3	71.4	3
UKA_20041123_D_Segment1	263.7	263.7	263.7	1
UKA_20041123_D_Segment2	25.2	4.1	14.6	2
UKA_20041123_E_Segment1	270.8	14.4	142.6	2
UKA_20041123_E_Segment2	640.2	45.4	342.8	2
UKA_20041124_A_Segment1	237.8	237.8	237.8	1
UKA_20041124_B_Segment1	247.9	247.9	247.9	1
UKA_20041124_B_Segment2	216.2	47.6	96.2	4
UKA_20050112_Segment1	283.9	283.9	283.9	1
UKA_20050112_Segment2	284.0	7.2	102.8	3
UKA_20050126_Segment1	245.6	245.6	245.6	1
UKA_20050127_Segment1	302.2	302.2	302.2	1
UKA_20050128_Segment1	290.7	290.7	290.7	1
UKA_20050128_Segment2	215.3	4.3	72.0	5
UKA_20050202_Segment2	63.9	2.3	23.6	7
UKA_20050209_Segment1	165.6	165.6	165.6	1
UKA_20050209_Segment2	58.7	28.0	40.6	4
UKA_20050310_A_Segment1	288.5	288.5	288.5	1
UKA_20050310_A_Segment2	161.7	27.8	90.1	4
UKA_20050310_B_Segment1	292.1	292.1	292.1	1
UKA_20050314_Segment1	250.7	250.7	250.7	1
UKA_20050314_Segment2	62.0	0.9	19.1	4
UPC_20050601_Segment1	394.7	18.6	166.9	3
all	816.3	0.9	134.1	70

Table 5.7: Average total speech time per speaker in each RT-06S lecture development meeting (in second).

5.1.4 Speaker count analysis

It is interesting to examine the speaker count as this is a key issue of the speaker clustering stage. As presented in [Wooters *et al.*, 2004], the iterative segmentation and clustering algo-

rithm initializes the models by dividing the data into a number of classes larger than the correct speaker number. This initial cluster number has been found to be capable of affecting largely the performance of clustering and needs to be adapted to the different types of data.

Table 5.8 shows the minimum, maximum and average speaker number on both broadcast news and meetings data. In general, the broadcast news shows contains much more total speakers than meetings. The speaker number in BN shows reaches 17 in average, while it varies greatly between the different shows, from 3 for the "20031115_180413_CSPAN_ENG" show to 27 for the "20010206_1830_1900_ABC_WNT" and "20031120_003511_PBS_ENG" shows. This large variation requires that the speaker diarization systems can provide adequate results, not only when a large number of speakers occur, but also when very few speakers exist.

In the meeting domain, there is a small amount of speakers in both conference and lecture meetings. The average speaker number is very low for the lecture room meetings (only 2) because several UKA lecture excerpts comprise just one speaker, whereas the "UKA_20050202_Segment2" consists of 7 speakers as well. The conference meetings have a little more speakers compared with lectures, with an average of 4 speakers. Therefore, the speaker diarization systems for meetings are usually tuned to produce less final clusters.

dataset	*excerpts number*	*max. speaker number*	*min. speaker number*	*ave. speaker number*
BN dev04f	12	27	3	17
conf dev06s	10	9	4	5
lect dev06s	30	7	1	2

Table 5.8: Average speaker number for broadcast news, conference and lecture meetings datasets.

5.2 Modifications to BN diarization system for meetings

Following the previous analysis of the differences between the broadcast news and meetings data, the adaptations of the LIMSI multi-stage broadcast news diarization system (c.f. 4.2) to meeting data will be described in this section.

5.2.1 Log-likelihood ratio (LLR) based SAD

The initial results on the development data from the baseline BN diarization system produced a high speech activity detection error, especially with a lot of missed speech, thus a different SAD approach was explored. One weakness of the standard Viterbi decoding is the lack of temporal control for each model. A transition penalty can be used to control the size of the segments, but as the level of noise increases, the likelihood of the speech model will decrease and thus the shortest speech segments will be discarded. Instead of setting a minimal likelihood level for

switching from one model to the other, it is easier to choose a minimal duration for speech and non-speech segments.

A simple speech activity detector based on the log-likelihood ratio (LLR) was applied. This detector calculates the LLR between the speech and non-speech models for each frame of the audio, and replaced the Viterbi decoding with a simple smoothing of the LLR followed by a decision module. More precisely:

- for each frame x_i, the LLR r_i between the speech and non-speech models λ_S and λ_{NS} is computed taking into account their prior probabilities $P(S)$ and $P(NS)$:

$$r_i = \log f(x_i|\lambda_S)P(S) - \log f(x_i|\lambda_{NS})P(NS)$$

- two adjacent smoothing windows with a duration of w frames (typically set to 50 or 100, i.e. 0.5 or 1 second) sliding over the signal are used for the detection of speech and non-speech transitions. A transition is possible when the sign of the averaged LLR in the left and right windows changes around the current frame:

$$s_i^+ . s_i^- < 0 \quad \text{with} \quad s_i^+ = \frac{1}{w} \sum_{j=i+1}^{i+w} r_j \quad \text{and} \quad s_i^- = \frac{1}{w} \sum_{j=i-w}^{i-1} r_j$$

- for a set I consisting of contiguous candidate transitions, the position of the transition is chosen at the maximum of difference between the averaged ratio of the left and right windows:

$$i^* = \underset{i \in I}{\text{argmax}} \, |s_i^+ - s_i^-|$$

The GMMs for speech and non-speech were trained on about 2 hours of far-field data from UKA seminars recorded in 2003 that were used as the test data in CHIL 2004 evaluation.

5.2.2 Applying voicing factor to SAD

Normally, the LLR-based speech activity detector is performed on cepstral coefficients with their Δ and $\Delta\Delta$ plus Δ and $\Delta\Delta$ log-energy. The reason for not using energy directly is that its level is sensitive to recording conditions, and it needs to be carefully normalized. Experiments on broadcast news data had led us to discard this feature. In order to further improve the SAD performance, a new energy normalization method taking into account a voicing factor v along with the energy E_0 is proposed. For each frame, the voicing factor is computed as the maximum peak of the autocorrelation function (excluding lag zero). The harmonic energy is thus defined as the energy associated to the best harmonic configuration, i.e. $E_h = v.E_0$. Finally, the energy of the signal is normalized relative to a reference level determined on the 10% frames carrying

the highest harmonic energy. This way, the energy normalization focus primarily on the voiced frames and may be more robust to varying SNR configurations. This method may be sensitive to music, but it is not expected to be an issue in the context of conferences and lectures.

5.3 RT-06S experimental results on lectures

LIMSI participated in the speech activity detection and speaker diarization tasks of the RT-06S evaluation, focusing on the lecture data. The modified diarization system for lectures with the new LLR-based SAD module (c.f. Section 5.2.1) was tested on far-field conditions: the Multiple Distant Microphone (MDM), Single distant Microphone (SDM) and Multiple Mark III Microphone Array (MM3A). The experimental results on the RT-06S lecture room datasets will be given in this section. The configurations of BIC clustering and SID clustering were optimized on the development data. All experiments were carried out with the BIC penalty weight $\lambda = 3.5$ and the SID threshold $\delta = 0.5$.

5.3.1 Performance measures and databases description

The primary metric used to measure the speaker diarization task performance is the same one that has been employed in the evaluations on BN data. As described in Section 3.1.1, the overall diarization error rate (DER) includes the missed and false alarm speaker errors and the speaker match error. The SAD task performance is evaluated by summing the missed and false alarm speech errors, without taking into account different reference speakers. The missed speech error is calculated as the ratio of the time that is labeled as non-speech in the hypothesis but as speech in the reference to the total speech time. The false speech error is computed as the ratio of the time that is labeled as speech in the hypothesis but as non-speech in the reference to the total speech time.

The SAD error is normally included in the missed and false alarm speaker errors. The missed speaker error is composed of the missed speech error and the hypothesis speaker time that is identified as a reference speaker having no mapped hypothesis speaker; the false alarm speaker error contains not only the false alarm speech error but also the reference speaker time that is attributed to a hypothesis speaker without a mapped reference speaker. Although the primary metric used in RT-06S evaluation is calculated over all the speech including the overlapping speech, the DER restricted to non-overlapping speech segments is also given for comparison purposes.

The experiments were conducted on the NIST RT-06S evaluation data comprised of lectures provided by the CHIL consortium. The development dataset *dev06s* consists of all seminars used as RT-05s evaluation data, plus an additional seminar from UKA and four seminars from AIT, IBM, ITC and UPC one each, with the duration ranging from 69 seconds to 937 seconds in each. The evaluation dataset *eval06s* is composed of 38 lecture recordings each lasting about 5 minutes.

5.3.2 Audio input selection

For the MDM evaluation condition, a single microphone signal randomly chosen from the available MDM channels different from channel selected for the SDM condition was used as the input to the speaker diarization system. Because the same microphone type is used for the MDM and SDM conditions, no individual development was carried on SDM condition, i.e. the same configuration for the speaker diarization system is adopted for both conditions. The microphone channels used for the MDM and SDM conditions are detailed in Table 5.9. For MM3A evaluation condition, the beamformed multiple mark III microphone array data provided by UKA was used as the input of the speaker diarization system.

dataset	*condition*	*AIT*	*IBM*	*ITC*	*UKA*	*UPC*
lect dev06s	MDM	mic05	Audio_17	Table-1	TableTop-1	channel15
lect eval06s	MDM	mic06	Audio_17	Table-2	TableTop-1	channel16
lect eval06s	SDM	mic05	Audio_19	Table-1	Table-2	channel15

Table 5.9: Channel selection for the MDM and the SDM conditions for the dev and eval data.

5.3.3 Results with different SAD on RT-06S development data

The performances of the speaker diarization systems integrating different SAD modules are summarized in Table 5.10. The "vit-bn" system uses Viterbi decoding with 5 GMMs (64 Gaussians) for speech, noisy speech, speech over music, pure music, and silence, each trained on one hour of BN data. This baseline speaker diarization system is the same system as was used in RT-04F evaluation for BN data. The "vit-bn+mt" system uses Viterbi decoding with GMMs trained on the BN data plus 2 GMMs (256 Gaussians) for speech (S) and non-speech (NS) trained on 2 hours of far-field data from the UKA seminars. The "vit-mt" system uses Viterbi decoding only with speech and non-speech models trained on lecture data. The "slr-mt" system uses the smoothed LLR-based SAD method with a prior probability of 0.2 for the non-speech model and 0.8 for the speech model. As can be seen in Table 5.10, Viterbi SAD using the models trained on both BN and lecture data have very high missed speech error rates (ranging from 18% to 14%) on the MDM development data. The log-likelihood based SAD substantially reduces this error (2.7% missed speech error) with limited increase in false alarm speech error. Compared with the baseline speaker diarization system, a relative DER reduction of 33% is obtained by the system using the smoothed LLR-based SAD.

5.3.4 Models with different number of Gaussians in LLR-based SAD

Table 5.11 gives the speaker diarization results on the MDM development data when the number of Gaussians for the speech and the non-speech models used in the smoothed LLR-based SAD

system	*missed speaker error (%)*	*false alarm speaker error (%)*	*speaker match error (%)*	*overlap DER (%)*
vit-bn (baseline)	18.2	3.0	9.0	30.2
vit-bn+mt	19.3	2.9	8.7	31.0
vit-mt	14.2	3.7	12.4	30.2
slr-mt	**2.7**	**6.1**	**11.7**	**20.5**

Table 5.10: Speaker diarization errors on the MDM development data for different SAD modules ("vit-bn" represents that the SAD is performed by a Viterbi decoding using 5 BN GMMs, "vit-bn+mt" indicates the Viterbi SAD using the BN models plus S/NS models trained on lecture data, "vit-mt" means that the Viterbi decoding is carried out to detect speech activities with using only 2 lecture SAD models, "slr-mt" represents the smoothed LLR-based SAD detector with 2 lecture SAD models using the prior probabilities for the non-speech and speech models as 0.2:0.8).

are varied. These results are obtained with a prior probability of 0.4 for the non-speech model and 0.6 for the speech model. There are no gains of the overall diarization error when the number of Gaussians is increased from 256 to 512 on the MDM development data.

nb. Gaussians	*missed speaker error (%)*	*false alarm speaker error (%)*	*speaker match error (%)*	*overlap DER (%)*
64	9.5	4.0	11.0	24.4
128	9.5	3.7	11.0	24.2
256	**7.8**	**4.2**	**11.0**	**23.0**
512	7.7	4.2	11.1	23.0

Table 5.11: Results of varying the number of Gaussians for the speech and non-speech models on the MDM development data, where the prior probabilities for the non-speech and speech models were set to 0.4:0.6.

5.3.5 Varied prior probabilities for S/NS models in LLR-based SAD

The effect of the prior probabilities for the speech and non-speech models used in LLR-based SAD was studied as well. The results presented in Table 5.12 are obtained with 256-component GMMs used for each model. Because it is important for automatic speech transcription to reject the least amount of speech as possible, a higher prior probability for the speech model is preferred relative to the non-speech model. As shown in Table 5.12, using a prior probability of 0.2 for the non-speech model and 0.8 for the speech model provides the best results for both speech activity detection (8.8% SAD error) and speaker diarization (20.5% DER).

P(NS):P(S)	*missed speaker error (%)*	*false alarm speaker error (%)*	*speaker match error (%)*	*overlap DER (%)*
0.1:0.9	1.0	9.5	12.0	22.4
0.2:0.8	**2.7**	**6.1**	**11.7**	**20.5**
0.3:0.7	5.2	5.0	11.3	21.5
0.4:0.6	7.8	4.2	11.0	23.0

Table 5.12: Results obtained by using different prior probabilities for the speech and non-speech models on the MDM development data.

5.3.6 RT-06S MDM lecture development results

After the experiments on the MDM development data, the configuration of the log-likelihood based SAD system is optimized as: a prior probability of 0.2 for the non-speech model and 0.8 for the speech model with 256-component GMMs used for both models. The performance of the speaker diarization system using the LLR-based SAD module is presented in Table 5.13, where the result is given for the individual seminar having the corresponding reference released by NIST. As can be seen in Table 5.13, the average DER of 20.5% masks the large variation across seminars. Normally lower overall diarization error can be obtained on the seminars with only one speaker, but for "UKA_20041124_A_Segment2" seminar, a very high false alarm speech error of about 150% is produced by the LLR-based SAD module. After listening to the audio file, it is found that many speech segments are missing in the reference transcription, this may be because the speech signal was not recorded on the microphone channel chosen for the manual reference transcription.

Table 5.13 also shows the overall diarization error (c.f. *DER**) from a 'do-nothing' diarization system where the whole audio recording is considered as the speech from one speaker. This blind system gives an overall non-overlap DER about 24% on the RT-06S lecture development data, which is close to the result from the developed lecture diarization system. This may be due to the nature of this specific lecture dataset, that is, some excerpts consist of a single speaker making a presentation during the whole audio recording. The performance comparison between these two systems demonstrates that there is still large space to improve system performances in the lecture meetings domain, even if the developed lecture diarization system reduces the global diarization error by 15% relatively. In fact, most error reduction come mainly from the false alarm speaker error. For the excerpts "UKA_20050126_Segment1", the developed system combining the LLR-based SAD module achieves a relative DER reduction of 83% compared to the blind system, with the false alarm speaker error decreased from 18.4% to 2.9%, as the overall DER from the blind system is equal to the FA error when there is only one speaker in the lecture.

In order to analyze the variation in system performance, the ratio between the speech time from the main speaker (who spoke the most in the seminar) and the total seminar duration is calculated for all the seminars in Table 5.13 except the "UKA_20041124_A_Segment2" seminar. Figure 5.2 shows that the speaker diarization system provides lower overall diarization error on seminars

meeting	*MS (%)*	*FA (%)*	*SPE (%)*	*DER (%)*	*DER* (%)*	*#REF*
AIT_20050726_Segment1	0.9	11.3	10.7	22.9	23.2	4
IBM_20050824_Segment1	2.6	0.9	1.6	5.1	6.3	2
ITC_20050429_Segment1	2.3	4.1	8.1	14.6	11.0	3
UKA_20041123_A_Segment1	0.0	0.9	0.0	0.9	3.6	1
UKA_20041123_A_Segment2	0.9	0.0	29.6	30.6	31.1	2
UKA_20041123_B_Segment2	7.2	35.0	4.6	46.7	76.6	3
UKA_20041123_C_Segment1	0.6	1.6	0.0	2.2	5.3	1
UKA_20041123_C_Segment2	1.7	5.7	4.1	11.4	21.1	3
UKA_20041123_D_Segment1	7.9	1.3	0.8	10.1	6.0	1
UKA_20041123_D_Segment2	1.6	75.5	4.8	81.9	128.8	2
UKA_20041123_E_Segment1	1.4	0.8	7.6	9.8	12.0	2
UKA_20041123_E_Segment2	3.5	5.1	6.6	15.2	17.1	2
UKA_20041124_A_Segment1	1.7	7.6	0.1	9.4	22.1	1
UKA_20041124_A_Segment2	3.2	149.9	4.5	157.6	174.9	1
UKA_20041124_B_Segment1	0.2	1.4	0.3	1.8	13.4	1
UKA_20041124_B_Segment2	1.9	3.2	44.2	49.3	56.6	4
UKA_20050112_Segment1	4.5	0.3	0.0	4.8	2.5	1
UKA_20050112_Segment2	10.2	1.2	7.2	18.7	13.7	3
UKA_20050126_Segment1	0.3	2.9	0.0	3.2	18.4	1
UKA_20050127_Segment1	1.3	0.2	0.1	1.6	0.9	1
UKA_20050128_Segment1	2.3	1.1	0.0	3.5	1.5	1
UKA_20050128_Segment2	3.0	1.7	39.5	44.1	44.6	5
UKA_20050202_Segment2	8.8	11.0	57.4	77.2	114.5	7
UKA_20050209_Segment1	2.5	1.4	0.0	3.9	11.8	1
UKA_20050209_Segment2	11.4	11.8	52.7	75.9	78.9	4
UKA_20050310_A_Segment1	0.5	1.7	0.8	3.0	7.1	1
UKA_20050310_A_Segment2	1.0	4.1	53.9	59.0	60.7	4
UKA_20050310_B_Segment1	0.2	0.6	0.0	0.8	3.1	1
UKA_20050314_Segment1	3.1	1.8	0.5	5.4	15.1	1
UKA_20050314_Segment2	6.0	3.3	19.0	28.3	49.1	4
UPC_20050601_Segment1	2.7	24.4	20.0	47.1	57.5	3
all	2.7	6.1	11.7	20.5	24.2	71

Table 5.13: Results by seminar in the MDM development dataset (*MS* is missed speaker error, *FA* is false alarm speaker error, *SPE* is speaker match error, *#REF* represents the speaker number in the reference), where *DER* indicates the diarization error generated by the lecture diarization system incorporating the LLR-based SAD module, *DER** presents the blind diarization result by hypothesizing that the entire signal is speech from one speaker.

where the main speaker spoke for more than 80% of the seminar duration. Moreover a correlation between the DER and the dominant speaker duration ratio is apparent clearly; consistent with the observations reported in [Mirghafori and Wooters, 2006].

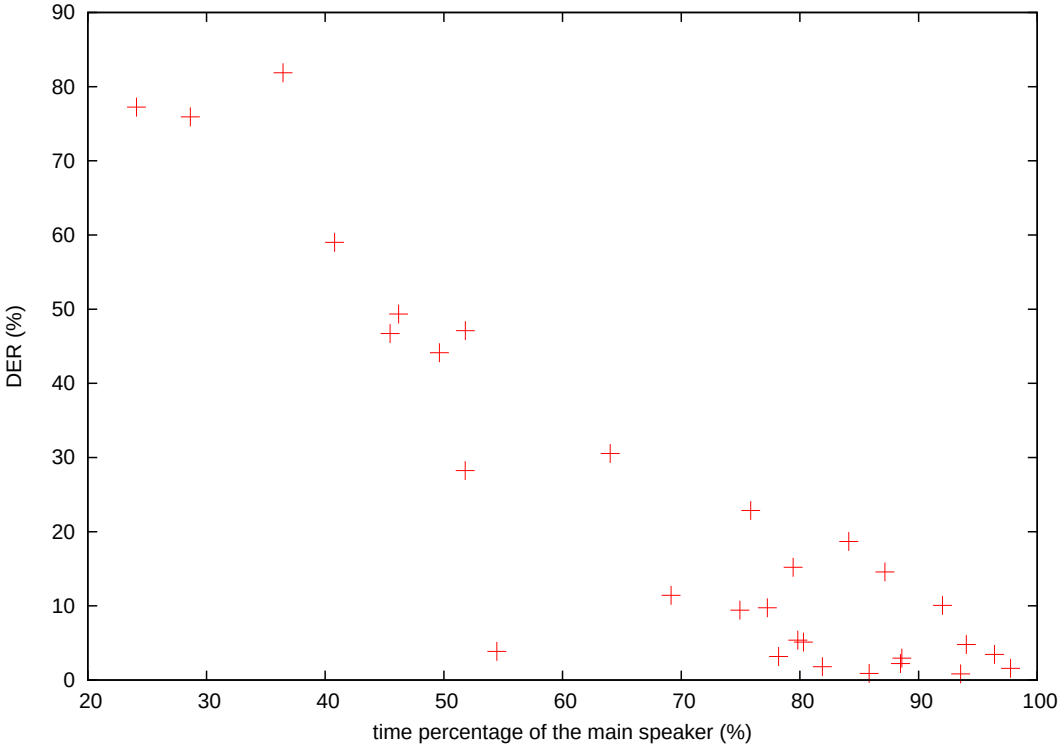


Figure 5.2: Overall speaker diarization error on the MDM development data as a function percentage of time of speech from the main speaker and the seminar duration.

5.3.7 RT-06S evaluation results

The RT-06S evaluation results are given in Table 5.14. For the MDM and SDM conditions, system tuning used the same development data, and therefore identical configurations are used for both conditions. The system performance is quite similar to that obtained on the MDM development data with an overlap overall diarization error of 21.5%. For the SDM audio input condition, the overlap DER is increased to 24.5%. This increase of the diarization error comes mainly from the SAD error, due to the different quality of the microphone channels used for the MDM and SDM conditions. In [Fiscus *et al.*, 2006]

For the MM3A contrast condition, the system configuration was optimized on the beamformed development data. Since no adaptation of the SAD acoustic models is performed on the beamformed data, a slightly higher diarization error of 25.9% is obtained for the MM3A condition

relative to the MDM condition.

condition	*nb. Gaussians*	*P(S)*	*overlap SAD error (%)*	*overlap DER (%)*	*non-overlap DER (%)*
MDM	256	0.8	9.0	21.5	20.2
SDM	256	0.8	12.4	24.5	23.2
MM3A	128	0.6	11.5	25.9	24.7

Table 5.14: Evaluation results for SAD and speaker diarization for the MDM, SDM and MM3A lecture conditions.

The RT-06S evaluation results from all participants can be found in [Fiscus *et al.*, 2006]. The ICSI diarization system gave the best performance on the lecture MDM test data (i.e. a overlap DER of 13.7%). This system useed a signal enhancement pre-process to generate a single signal from all available MDM channels and the output inter-channel delay features were further mixed with the MFCC acoustic features via replacing the standard log-likelihoods used in the BIC clustering with the combined log-likelihoods between both sets of features[Anguera *et al.*, 2006c; Pardo *et al.*, 2006a; Pardo *et al.*, 2006b]. The experiments reported in [Anguera, 2006] have shown that using forced alignment reference segmentations can provide a significant improvement of the SAD and diarization performances. The system developed at the university of Avignon had an overlap DER of 24.5% on the RT-06S lecture MDM test data, in which one single signal was obtained by simply combing all available channels with their weights computed relying on the SNR estimations [Fredouille and Senay, 2006].

5.4 RT-07S experimental results on conferences and lectures

Beside the conference room and lecture room sub-domains, a new type of recordings has been introduced into the NIST 2007 Spring meeting recognition evaluation [NIST, 2007], that is, the recordings of coffee breaks. Since the lecture room and coffee break excerpts were extracted from different parts of the same meetings, both sub-domains have the same sensor configurations but these are different from the conference meeting sub-domain. Although these three types of data have different styles of participant interaction, the RT-07S lecture evaluation data contains more interactions between meeting participants than the previous lecture meeting data. Some other changes from the RT-06S evaluation to the RT-07S evaluation are made in the aspect of the task definition, i.e. the speech activity detection task has been discarded and the "Speaker Attributed Speech-To-Text" (SASTT) has been introduced as a new task that combines both speaker diarization and speech-to-text technologies into a single joint task and provides a text output consisting of the spoken words each with the corresponding speaker.

In the RT-07S evaluation, LIMSI participated to the speaker diarization task on the conference and lecture data and developed a general diarization system for both types of meeting data. As the RT-06S lecture diarization system had a high SAD error that in turn affects strongly final

speaker diarization performance, some new acoustic models trained using the forced alignment transcriptions and several different feature normalization techniques are employed to reduce the speech activity detection error. In order to ameliorate the SID clustering performance, several new UBMs were also constructed in a similar manner. Since the RT07S evaluation plans specified the use the word-forced alignment reference segmentations for scoring speaker diarization performance, both the SAD acoustic models and UBM are retrained using forced alignment segmentations. The forced alignment transcriptions were derived from the manual transcriptions via the ICSI-SRI ASR system [Stolcke *et al.*, 2005] that aligns the spoken words to the signal, and therefore segment boundaries are labeled more accurately than the hand-made ones. Based on the aligned segmentations, more precise speech and non-speech models can be estimated and are expected to provide better SAD performances. Instead of training on the UKA lecture data that was used in the RT-06S diarization system, an union of several previous RT conference datasets consisting of recordings from different sites was used to better approximate the RT-07S test data.

For the MDM audio input condition, the RT-07S diarization system uses the beamformed signals generated from the ICSI delay&sum signal enhancement system [Anguera *et al.*, 2005a] instead of selecting randomly one channel from all available MDM channels as was done by LIMSI RT-06S lecture diarization system. Since the telephonic speech is assumed not to occur in meeting data, the bandwidth detection stage is removed from the baseline BN system and the SID clustering is performed separately within each gender class (male/female).

The development experiments and evaluation results within the framework of the RT-07S evaluation will be given in this section. The diarization performances are measured via the same metrics described in Section5.3.1. Unless otherwise specified, the development experiments were carried out with a BIC penalty weight $\lambda = 3.5$ and a SID threshold $\delta = 0.5$.

5.4.1 Database description

In order to tune the system parameters, the development experiments were carried out on the RT-06S test data in both the conference and lecture sub-domains. The conference development dataset *conf dev07s* is composed of 9 conference meetings, with a duration of about 15 minutes each. This is the same corpus that was used as the test data in the RT-06s evaluation and was provided by 5 different laboratories: CMU (Carnegie Mellon University), EDI (The University of Edinburgh), NIST, TNO (TNO Human Factor) and VT (Virginia Tech). The lecture development data includes two corpora: the RT-06S lecture evaluation dataset (denoted as *lect dev07s1*) and a new development dataset (denoted as *lect dev07s2*) released in 2007. The *lect dev07s1* consists of 38 5-minutes lecture excerpts contributed respectively by 5 of the CHIL partner sites: AIT, IBM, ITC, UKA and UPC, for which only 28 excerpts reference segmentations are available. The *lect dev07s2* contains 5 lectures with different audio lengths ranging from 23 minutes to 44 minutes, recorded more recently at the same 5 CHIL sites.

5.4.2 LLR-based SAD with different acoustic features

Although the speech activity detection task is not included in the RT-07S evaluation, a good SAD module is always useful for a speaker diarization system in the sense that it can influence the accuracy of the acoustic models which serve in the subsequent segmentation and clustering stages. Therefore, different kinds of acoustic features were investigated within the LLR-based SAD module. To do this, the SAD stage is separated from the speaker diarization system and assessed as an independent system on the RT-07S development data. However, an optimal SAD is not necessarily the best choice for diarization systems, as false alarm speech will corrupt the speaker models used in clustering stage: the experiments made at ICSI also show that minimizing short non-speech data is helpful to improve diarization performance [Wooters and Huijbregts, 2007].

Table 5.15 gives the SAD results with using different acoustic features, where each type of features has its appropriate SAD acoustic models estimated on the training data parameterized by the same features. In all cases, the speech and non-speech models are trained on the same data consisting of 8 RT-04S development conference meetings, 8 RT-04S evaluation conference meetings and 10 RT-05S evaluation conference meetings. It should be noted that the forced alignment segmentations provided by ICSI-SRI were used to estimate the SAD acoustic models. The lack of the forced alignments for the lecture data except the *lect dev07s1* (which serves as development data) is the reason of using only the conference meetings to construct the speech and non-speech models for both the conference and lecture test data.

acoustic features	*missed speech error (%)*	*false alarm speech error (%)*	*overlap SAD error (%)*
conf dev07s			
baseline	1.3	4.3	5.6
baseline+e	1.1	4.0	5.1
baseline+evn	1.1	3.3	4.3
baseline+e+mvn	0.8	3.0	3.9
lect dev07s1			
baseline	2.4	5.3	7.8
baseline+e	0.5	11.2	11.8
baseline+evn	0.9	4.7	5.7
baseline+e+mvn	1.0	5.6	6.6

Table 5.15: SAD results obtained with using different kinds of acoustic features on the RT-07S beamformed MDM development data.

The SAD results shown in Table 5.15 are obtained with the same LLR-based SAD configurations: the smoothing window with a size of 50 frames (i.e. 0.5 sec) and the prior probabilities respect to the non-speech and speech models being $0.2 : 0.8$ with 256 Gaussians components in

each model. The baseline vector consists of 12 cepstral coefficients with their Δ and $\Delta\Delta$ plus Δ and $\Delta\Delta$ log-energy and provides a SAD error of 5.6% and 7.8% on the conference and lecture development datasets respectively. When the raw energy is simply appended into the baseline acoustic vectors (c.f. denoted as "baseline+e" in Table 5.15), the SAD error is reduced to 5.1% for the conference dataset but increases to 11.5% for the lecture data. This degradation of the SAD performance on the lectures is predictable due to the variability of recording conditions in different lecture rooms. The SAD error reduction obtained on the conference meetings suggests that the recording conditions are consistent across the conference audio data. Replacing the energy with the normalized energy relying on the voicing factor v (denoted as "baseline+env") decreases the SAD error to 4.3% on the conference meeting data and 5.7% on the lecture meeting data. Regarding some details, the performance improvement obtained on the conferences comes from the reduction of the false alarm speech error, while the gain observed on the lectures derives merely from the missed speech error. The "baseline+e+mvn" feature set performs a variance normalization of each baseline feature and energy by subtracting their means and scaling by their standard deviations. Using this normalized acoustic representation, a further SAD reduction of absolutely 0.4% is obtained on the conference development data, but no improvement is obtained on the lecture dataset. In consideration of the simplicity of the diarization system, the SAD models trained with the "baseline+e+mvn" feature set were used for both the RT-07S conference and lecture evaluation data. Speech and non-speech models with 128 Gaussians were also investigated and although they gave a higher SAD error compared with 256 Gaussians, they perform slightly better on the diarization task.

5.4.3 SID clustering with UBMs trained on different acoustic features

The different sorts of acoustic features are also tested for the UBM training. A single gender-independent UBM with 128 Gaussian mixtures is trained for each type of feature, since no gender information is available in the forced alignment segmentations of the training data. Table 5.16 shows the speaker diarization results on the beamformed MDM development data using different acoustic features to train UBMs. These results are obtained with the same SAD acoustic models that were trained on the normalized features via the variance normalization technique. All UBMs are trained on the same conference dataset that have been used to estimate the speech and non-speech models.

The baseline feature vector is composed of 15 Mel frequency cepstral coefficients plus the Δ coefficients and the Δ log-energy with the feature warping normalization (referred to as "15plp+Δ+ΔlogE+w" in Table 5.16). This baseline feature set provides an overlap DER of 36.2% on the conference development data and 17.5% on the lecture development data. Adding the $\Delta\Delta$ coefficients and the $\Delta\Delta$ log-energy, namely "15plp+Δ+$\Delta\Delta$+ΔlogE+$\Delta\Delta$logE+w" reduces the DER to 31.1% on the conferences but gives a similar diarization performance as the baseline features for the lectures. When the dimension of cepstral coefficients is diminished to 12, using the "12plpΔ+ΔlogE+w" feature set results in a further DER reduction of 0.5% on the conference data but no significant performance change on the lecture meetings. Appending the

acoustic features	*speaker match error (%)*	*overlap DER (%)*
conf dev07s		
15plp+w*	20.5	28.3
15plp+Δ+ΔlogE+w	28.4	36.2
15plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	23.3	31.1
12plp+w*	21.3	29.0
12plp+Δ+ΔlogE+w	22.9	30.6
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	27.9	35.7
12plp+mvn*	28.6	36.3
12plp+Δ+ΔlogE+mvn	33.8	41.6
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+mvn	32.0	39.8
lect dev07s1		
15plp+w*	10.3	17.8
15plp+Δ+ΔlogE+w	10.0	17.5
15plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	10.2	17.7
12plp+w*	11.3	18.8
12plp+Δ+ΔlogE+w	10.3	17.8
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w	10.2	17.7
12plp+mvn*	10.1	17.6
12plp+Δ+ΔlogE+mvn	10.5	18.0
12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+mvn	10.2	17.7

Table 5.16: Diarization results obtained with the UBMs trained on different acoustic representations on the RT-07S beamformed MDM development data (results with '*' were obtained after the evaluation).

ΔΔ coefficients (denoted as "12plp+Δ+ΔΔ+ΔlogE+ΔΔlogE+w"), a large increase of DER is observed on the conference data but the DER score rests always very close to the baseline one for the lectures. Finally, the variance normalization method was also examined within the SID clustering stage. As can be seen in Table 5.16, higher DER rates are provided by this normalization approach than the feature warping technique on both the conference and the lecture development data. For the lecture data, the different acoustic representations are found to give similar diarization results. This may be derived from the mismatch between the conference training data and the lecture test data.

The results of the post-evaluation experiments without the use of derivative parameters are also given in Table 5.16. Using only static features reduces the diarization error rates on the conference development data and the lowest DER of 28.3% is given by the 15 MFCC with the feature warping normalization. However, no significant difference in diarization performances is observed on the lecture data between using only the static coefficients and by appending the Δ and ΔΔ coefficients.

5.4.4 RT-07S evaluation results

The speaker diarization results for the systems submitted to the RT-07S evaluation are given in Table 5.17. The diarization system uses the same SAD acoustic models and UBM trained on the "baseline+e+mvn" and the "12plp+Δ+ΔlogE+w" feature sets respectively for both the conference and the lecture evaluation data. The BIC penalty weight λ and the SID clustering threshold δ were optimized on the development data and set individually for the conference and lecture test data: $\lambda = 3.5, \delta = 0.6$ for the conference dataset and $\lambda = 3.5, \delta = 0.5$ for the lecture dataset. For each type of the data, the same system configurations were used on the MDM and SDM audio input conditions.

For the conference test data, the diarization system has an overall diarization error of 26.1% on the beamformed MDM signals, and the overall DER increases to 29.5% for the SDM condition. The beamformed signals from all available distant microphones are shown to be helpful for improving the diarization performance. For the lecture evaluation data, the diarization system gives similar performances on both the beamformed MDM signals and a single SDM data. This may be because the delay&sum signal enhancement system was not well tuned for lecture data. A much higher false alarm error is observed on the lecture test data than on the conference due to the mismatch between the SAD training data and the lecture test data.

data type & condition	*MS (%)*	*FA (%)*	*SPE (%)*	*overlap DER (%)*	*non-overlap DER (%)*
conference MDM	4.5	1.3	20.2	**26.1**	23.0
conference SDM	4.9	1.3	23.3	**29.5**	26.6
lecture MDM	2.6	8.4	14.7	**25.8**	24.5
lecture SDM	2.9	8.1	14.7	**25.6**	24.3

Table 5.17: Speaker diarization performances on the RT-07S conference and lecture evaluation data for the MDM and SDM conditions.

A 'do-nothing' diarization system that assumes only one speaker occurs in each meeting excerpt was also carried out on the lecture test dataset. This blind system gives an overall DER of approximately 30% on the beamformed MDM signals. The LIMSI meeting diarization system reduces the overall diarization error to 25.8%, thus providing a relative DER reduction of 13%. In addition, the effect of the SID clustering threshold δ on the speaker match error and the cluster purity error was also investigated by the post-evaluation experiment on the lecture MDM data. As shown in Figure 5.3, the speaker match error keeps relatively lower when the SID threshold varies between 0.2 and 0.5, even if the minimum error is obtained with δ set to around 0.2. The speaker match error decreases with no significant degradation of the cluster purity when the SID threshold is reduced from 2.0 to 0.5. This threshold effect is similar as that are obtained on BN data (c.f. Figure 4.6 and 4.8).

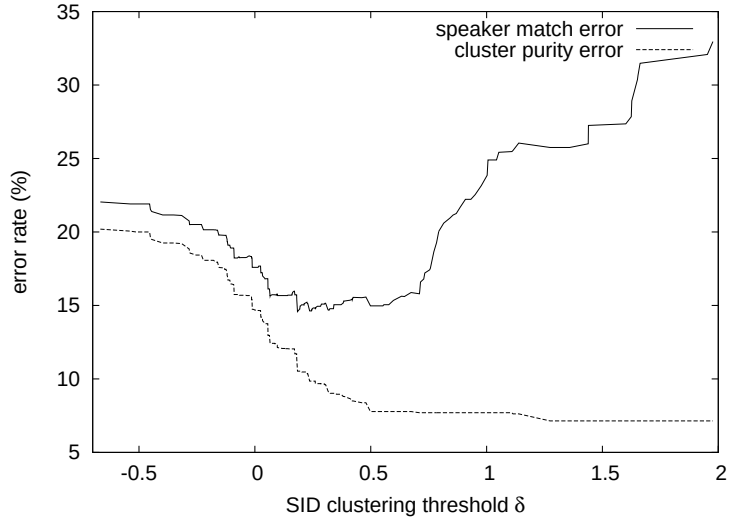


Figure 5.3: Speaker match error and purity error rates on the RT-07S lecture beamformed MDM data as a function of the SID clustering threshold δ.

5.5 Conclusions

The LIMSI RT-06S speaker diarization system for the lecture data builds upon the baseline multi-stage system developed for the broadcast news. The main modification is the use of the log-likelihood based SAD with the acoustic models adapted on the lecture data. This LLR-based SAD has been demonstrated to perform much better than Viterbi SAD which is used in the baseline system. On the MDM development data, the LLR-based SAD provides a significant reduction on the SAD error up to 58% relative to Viterbi SAD, especially on the missed speech error. Concerning the speaker diarization performance, the diarization system using the LLR-based SAD gives an overall error of 20% , while a 30% overall error is obtained by the baseline system. On the evaluation data, the RT-06S speaker diarization system provides an overlap overall diarization error of 21.5% on the MDM condition, with a small increase in the overlap DER to 24.5% for the SDM condition and a higher error of 25.9% for the MM3A condition. The robustness of the speaker diarization system depends a lot on the data domain. The combination of BIC clustering and SID clustering is very effective on the BN data and provides 8.5% non-overlapping overall diarization error on RT-04F evaluation data. A relatively higher non-overlapping DER of 20.2% is obtaind on the MDM lecture data. This decrease of the speaker diarization performance may derive from the lower signal quality of the lecture data.

The RT-07S diarization system for both conference and lecture meetings keeps the main structure of the RT-06S lecture diarization system except removing the bandwidth detection module, since it is supposed that no telephonic speech would occur during the meetings. The main improvements come from the new SAD acoustic models and UBM that were built on the conference training data with their forced alignment segmentations. Using the speech and non-speech mod-

els trained with the variance normalized acoustic features yields the best SAD performance on the development data, i.e. 3.9% for the conference data and 6.6% for the lecture data. The mismatch between the conference training data and the lecture test data results in a relatively higher SAD error on the lecture development data. As for UBMs, the best diarization performance is generated by using the gender-independent UBM trained with 12 Mel frequency cepstral coefficients plus Δ coefficients and Δ log-energy with the feature warping technique. The adapted diarization system provides similar diarization results on the beamformed MDM signals for both the RT-07S conference and lecture evaluation data (i.e. an overlap DER of 26.1% for the conference dataset and 25.8% for the lecture dataset). The DER rate increases to 29.5% on the conference SDM data, while for the lecture SDM data, the error rate remains very close to the one obtained on the beamformed MDM condition.

Chapter 6

Overlapping speech detection

Overlapping speech is usually ignored by speaker diarization systems. This may be an acceptable choice for broadcast news recordings where the conditions are controlled, but in more spontaneous data the amount of overlapping speech increases. The focus of this chapter is the detection of overlapping speech in monaural signals. Experiments are carried out on telephone conversations for which the two separate channel recordings are available and transcribed. This allows the ground truth of overlapping speech to be easily determined when faced with the mixed signal. Each class, i.e. single speaker and overlapping speech, is modeled by a Gaussian mixture models (GMM) trained on acoustic features. Performance is assessed through the equal error rate (EER) of the frame-level overlapping speech detection. Standard cepstral and autocorrelation features are compared; their combination providing a 33.6% EER performance when the decision was smoothed over a 500ms window duration.

6.1 Introduction

Overlapping speech has since long been a subject of interest in the field of dialogue and conversation analysis [Sacks *et al.*, 1974; Levinson, 1983]. However, if the more specific issue of echo cancellation is excluded, only recently have research in automatic speech processing addressed the problem of speech overlaps. There may be two main reasons for this situation. First, due to the complexity of speech processing, automatic speech recognition (ASR) long focused on the tasks (dictation, broadcast news or human-machine communications) where overlapping speech was absent or rare. Second, overlapping speech was simply beyond the skills of ASR systems, and it remained possible to ignore such infrequent situations even if they induced locally a higher error rate, so speech segments detected in the audio signal were assumed to be spoken by a single speaker.

The situation has evolved, with more attention paid to spontaneous speech from multi-party conversations, especially telephone conversations and meetings. A study on ICSI meetings by

Shriberg *et al* [Shriberg *et al.*, 2001] reported that 6 to 14% of words are overlapped, even when simple backchannels are ignored, resulting in a large increase in transcription errors (these were later more deeply investigated by Çetin and Shriberg [Çetin and Shriberg, 2006]); on telephone conversations they reported about 8% of overlapped words. Also in telephone conversations, ten Bosh *et al* [Ten Bosch *et al.*, 2005] observed that overlapping speech occurred at 52% of changes in speaker turns. In a more controlled situation, the figure is of course lower, e.g. it was measured at 3% of foreground words in a corpus of TV political interviews [Adda *et al.*, 2007]. Given that ASR performance on broadcast news data has improved and can be as low as about 10% word error for the best systems, the impact of overlapping speech is no longer negligible.

In some situations, one can rely on multiple channel recording, allowing source separation or source localization. There have been a significant number of studies in this area. Pfau *et al* have proposed a multispeaker speech activity detection method for meetings [Pfau *et al.*, 2001]. Wrigley *et al* [Wrigley *et al.*, 2005] explored the use of different acoustic features for crosstalk detection in multi-channel audio data from the ICSI meeting corpus; finding the cross-channel correlation to be the most discriminant among the tested features. Laskowski and Schulz trained models for overlapped speech for improving the speech detection in meetings [Laskowski and Schultz, 2006]. The inter-channel time differences were found to be useful for speaker diarization in meetings [Pardo *et al.*, 2006b] and they can be used as a cue for multiple simultaneous speakers [Van Leeuwen and Konecny, 2007].

In a more general case where a multi-channel recording is not available, the detection of overlapping speech on a monaural signal is also of clear interest. This problem can be seen in the larger context of computational auditory scene analysis [Bregman, 1994; Cooke and Ellis, 2001; Wang and Brown, 2006], where the aim is to segregate the sound in the time-frequency domain according to the different sources. Speech separation of monaural signals is a difficult task. However, given the harmonic nature of voiced speech, the detection of overlapping speech is highly correlated with the multi-pitch detection and tracking problem [De Cheveigné, 2006]. It can be expected that approaches and features which are relevant for speech separation and multi-pitch detection are also of interest for speech overlap detection.

Since the detection approaches mentioned above can not provide satisfied detection performance, we try to propose a new algorithm for overlapping speech detection in a relatively simple case, i.e. conversational telephone speech (CTS). The reason for selecting CTS data is that there are usually only two speakers presenting in a telephone conversation and we have at our disposal databases where each speaker is recorded in a separate channel. This allows to train a model for overlapping speech that is obtained by mixing the speech signals from both speakers. It will be possible to adapt the proposed overlapping speech detection algorithm towards the meeting domain if it provides reasonable detection performance on CTS data.

This chapter aims to address overlapping speech detection in a two-speaker conversation with an a priori knowledge of the two speaker voices in isolation. We present experiments on telephone conversations for which the two separate channel recordings are available and transcribed. The Gaussian mixture models respectively for non-overlapping and overlapping speech classes are first trained on a part of conversation data. The log-likelihood ratio between two classes is

then computed for each frame of the remaining data based on cepstral and autocorrelation features individually. Performance was assessed using the equal error rate (EER) of the frame-level overlapping speech detection.

The next section describes the log-likelihood ratio based overlapping speech detection framework where the cepstral and autocorrelation features are investigated respectively, and some preliminary experimental results are given in Section 6.3.

6.2 LLR-based overlapping speech detection

The proposed overlapping speech detection method is based on the assumption that the training data of both participating speakers in a telephone conversation is available as well as the corresponding reference segmentation. In this paper, this a priori information is acquired by using the first half of a conversation as the training data and the remainder as the test data. These experiments use conversational telephone speech where each conversation side is recorded in a separate channel. The reference transcriptions of both conversation sides were obtained by aligning the manual reference transcription provided by LDC to the matching signal using the LIMSI transcription system [Gauvain *et al.*, 2003]. Then, the reference speaker segmentation was generated via combining the forced alignments of the two sides, which provides speech segments from each speaker and overlapping speech segments.

The basic idea is to identify overlapping speech regions using the models trained from the reference speaker segmentation. Assuming $C = \{(x_t, y_t)\}_{1 \leq t \leq T}$ is the sample sequence for a call, with x_t and y_t the sample for each of the two channels. Let $z_t = x_t + y_t, t \leq T$ be the mixed signal of both channels with the same mixture weight for each channel. Furthermore, let $A = \{x_{a_i}\}_{i \leq n_a}$ and $B = \{y_{b_j}\}_{j \leq n_b}$ be the set of speech samples associated to the speakers A and B from x_t and y_t sequence respectively, where the overlap speech samples are excluded for both speakers.

In the training phase, a model for the speech from only one speaker, namely non-overlapping speech model and an overlapping speech model are built on the first half data form the conversation. The non-overlapping speech model is trained on the concatenation of all speech segments determined during the alignment step from either speaker A or B determined extracted from the mixed signal:

$$\lambda_{no} \text{ on } \{z_{a_i}\}_{a_i \leq T/2} \cup \{z_{b_j}\}_{b_j \leq T/2}$$

Since each experimental conversation lasts approximately 5 minutes with about 5.5% of overlapping speech, the amount of the training data for overlapping speech is too small to estimate a robust model. We decided to artificially increase the amount of overlapping speech training data by combining the speech segments from both speakers extracted from the individual channels into one signal. The overlapping speech model is estimated from this mixed signal:

$$\lambda_o \text{ on } \{x_{a_k} + y_{b_k}\} \qquad a_k \leq T/2, b_k \leq T/2$$

The overlapping speech detection is performed on the second half of the mixed signal z_t. The log-likelihood ratio (LLR) between the non-overlapping and overlapping speech models is computed for each frame:

$$llr(t) = \log L(z_t|\lambda_o) - \log L(z_t|\lambda_{no}) \quad T/2 < t \leq T$$

A frame is identified as overlapping speech if the value of $llr(t)$ is greater than a specific threshold δ that is determined a posteriori to provide an equal error rate (EER) of the misclassification between overlapping and non-overlapping speech classes.

6.2.1 Using Mel frequency cepstral parameters

The proposed LLR-based overlapping speech detection technique is first carried out with the Mel frequency cepstral parameters that were used in a standard transcription system [Gauvain *et al.*, 2003]. The MFCCs are extracted from the signal every 10ms using a 30ms window on a 0-3800Hz band and the 39 dimensional feature vector consists of 12 cepstral coefficients, Δ and $\Delta\Delta$ coefficients plus the energy with Δ and $\Delta\Delta$ log-energy. This is similar to the feature set that is used in the baseline diarization system for broadcast news 4.1, except for the energy and extracting from different pass bands. A variance normalization technique is also performed on the MFCC coefficients. Gaussian mixture models (GMM) with 64 Gaussians and diagonal covariance matrices were trained on these coefficients and used for computing the frame-based LLR.

6.2.2 Using autocorrelation features

The cepstral features were introduced for speech transcription and carry acoustic information on the timbre of the phonemes and at a larger time scale, of the speaker. They have thus been successfully used for speaker recognition applications. However, part of the speaker-specific information also lies in the prosody, and the pitch information is normally suppressed or at least highly reduced by the cepstral features extraction. As an alternative to MFCC features, we also investigate a sampled short-term autocorrelation of the signal within the LLR-based overlapping speech detection framework. The autocorrelation is often used for voicing detection and pitch extraction:

$$r_t(\tau) = \sum_{j=t}^{t+W-1} x'_t x'_{t+\tau}$$

where W is the size of the window, τ the lag (homogeneous to a time delay) and x'_t a windowed version of the signal. We use the normalized autocorrelation $r'_t(\tau) = r_t(\tau)/r_t(0)$ so that $r'_t(0) = 1$. Next, the autocorrelation is sampled on a limited number of K bins. The frequency domain is sampled between F_{min} and F_{max} on a log scale:

$$F_k = F_{min} \left(\frac{F_{max}}{F_{min}} \right)^{\frac{k}{K}}, k \leq K$$

Finally, the R_k feature is taken as the maximum of the autocorrelation function for the lags corresponding to the $[F_{k-1}, F_k]$ frequency interval:

$$R_k = \max_{F_{k-1} < 1/\tau \leq F_k} r'_t(\tau)$$

R_k will show values near to 1 in the frequency range of the pitch, and the distribution of the coefficients should thus be related to the speaker pitch characteristics. In the case of overlapping speech, the maxima of the autocorrelation function for the overlapping signal reflects a distribution characterizing simultaneously the pitch of both speakers.

The default settings of the autocorrelation features computation are given as the following: the autocorrelation coefficients are computed over a 40ms analysis window at a 10ms frame shift, with $F_{min} = 75$Hz, $F_{max} = 600$Hz and $K = 16$. These features are further processed exactly the same way as the MFCC features for GMM training, including the Δ and $\Delta\Delta$ dynamic features plus energy and Δ and $\Delta\Delta$ log-energy, except no mean and variance normalization is performed.

6.2.3 Combining LLRs based on MFCC and autocorrelation features

To combine the two kinds of scores, the mean of the likelihood ratios obtained from both MFCC and autocorrelation features is computed for each frame. For the three types of scores, a smoothing procedure is performed via computing an average llh value over a given window size.

6.3 Experimental results

The experiments were conducted on a subset of 50 calls from Switchboard cellular part 1 corpus distributed by LDC [Graff *et al.*, 2001]. This data subset is composed of 25 conversations with matching gender speakers (9 male-male and 16 female-female conversations) and 25 conversations with cross-gender speakers. Each conversation lasts about 5 minutes and in principle includes two speakers. In this experiment dataset, 5.5% of the conversation data is the ovelapping speech. In this section, effect on overlapping speech detection of using different sampling numbers of the autocorrelation features (i.e. varying K in autocorrelation computation) and different

autocorrelation window sizes (i.e. changing the value of W) is studied on the examined dataset. The overlap detection results from using different feature sets are also reported in this section.

In order to measure the performance of overlapping speech detection, the reference speaker segmentation was converted into a frame-level segmentation. The detection error is measured only on the non-overlapping speech and overlapping speech regions and is expressed in terms of equal error rate (EER) that is determined by tuning the decision threshold to generate the equal missed and false alarm detection error rates. The detection error trade-off (DET) curve is also used to assess the performance, where the missed error rate is plotted as a function of the false alarm error rate.

6.3.1 Sampling of the autocorrelation features

The autocorrelation features with different scale numbers of the output coefficients (i.e. using different values of K) are investigated within the proposed overlapping speech detection scheme. The resulting equal error rates are demonstrated in Figure 6.1, where all the other parameters in the autocorrelation feature computation remain the default values. It is shown that the detection error decreases as the number of the output autocorrelation features is reduced. The lowest EER around of 34% is obtained using 10 output autocorrelation features with the size of the smoothing window set to 50 frames (i.e. 0.5 seconds). In all the cases, the EER curves display the same tendency of the overlapping speech detection performance and the best result is always achieved by using a 0.5s window to smooth log-likelihood ratios.

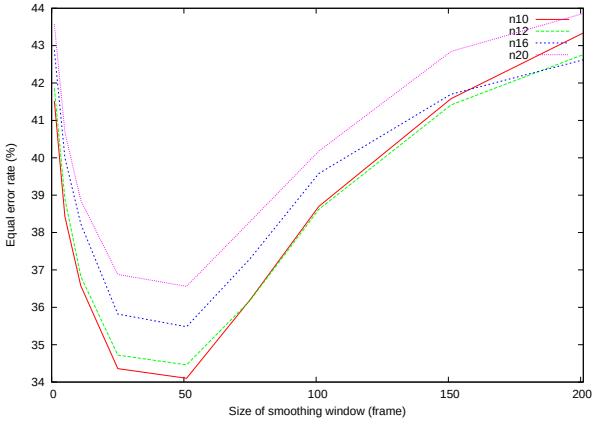


Figure 6.1: Equal error rate for overlapping speech detection obtained using different numbers of the output autocorrelation features as a function of the smoothing window size.

6.3.2 Autocorrelation window size

After the optimal number of autocorrelation features is tuned to 10, the LLR-based overlapping speech detection is performed with the autocorrelation features computed over the analysis windows of various lengths. Figure 6.2 shows the EERs of the overlapping speech detection results. It is found that lower equal err rate is obtained by using smaller analysis window to extract autocorrelation features. Using an analysis window of 30ms gives the best detection performance of 33.6% EER. Since the computation window is constrained to be larger than $2/F_{min}$, the minimum size of the window is 27ms with the default value of F_{min} set to 75.

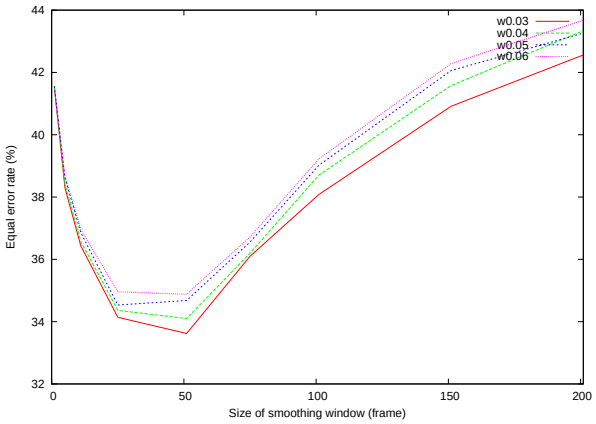


Figure 6.2: Equal error rate for overlapping speech detection obtained using different sizes of autocorrelation analysis window as a function of the smoothing window size.

6.3.3 DET curves for different feature sets

Figure 6.3 shows the DET curves from the LLR scores obtained using MFCC and autocorrelation features, as well as their combination. The three scores are smoothed over a 0.5 second window. The autocorrelation features are computed over a 30ms window at a stepping of 10ms, with $F_{min} = 75$Hz, $F_{max} = 600$Hz and $K = 10$. The detection system based on autocorrelation features provides better overlapping speech performance compared with the MFCC system. However, no significant performance improvement is obtained by combining the scores from using MFCC and autocorrelation features.

6.3.4 EER obtained on different gender combinations

Following the experimental results presented above, the equal error rates are also given in Table 6.1 for different gender combination conditions of the conversations (i.e. male vs. male, female vs. female and male vs. female). The overlapping speech detection system based on cepstral features (MFCC) has higher EERs for each gender combination condition compared with

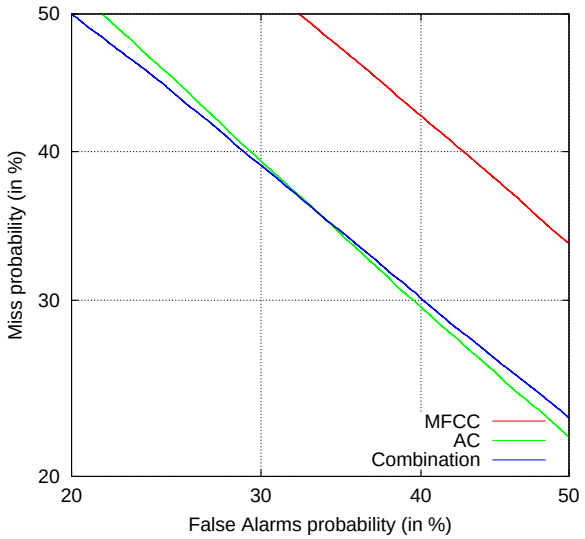


Figure 6.3: Detection error trade-off (DET) curves for overlapping speech detection based on cepstral (MFCC), autocorrelation (AC) features and their combination.

the systems relying on autocorrelation features and combined scores. The best detection performance (i.e. 31% EER) is observed on the conversations between two female speakers via the system based on autocorrelation features, with an EER of about 33% obtained by using combined scores from MFCC and autocorrelation coefficients. The conversations between male speakers seem to be more difficult for overlapping speech detection, as the EER approximately of 35% is provided by the system using autocorrelation features and combined LLR scores. The conversations between cross-gender speakers are found not to be easier for overlapping speech detection than ones between same gender speakers and the detection performance on them is in between the performances for the female vs. female and male vs. male conversations.

Score type	*M-M*	*F-F*	*M-F*
MFCC	42.4%	42.1%	40.4%
AC	35.0%	31.3%	34.5%
combination	35.4%	32.8%	34.9%

Table 6.1: Equal error rate for overlapping speech detection obtained on the data with different gender combination conditions.

6.4 Conclusions

A log-likelihood ratio based overlapping speech detection approach is introduced in this chapter. Experiments were conducted on telephone conversational speech, where the first half of each

conversation is used to train the GMMs for non-overlapping and overlapping speech classes respectively and the rest is reserved for the test data. In this detection framework, we investigated cepstral and autocorrelation features as well as their combination. The primary experimental results show that the overlapping speech detection system based on autocorrelation features outperforms the one replying on Mel frequency cepstral coefficients. The combined system has a similarly detection performance as the autocorrelation feature system and both systems give the minimum EER of 33.6% on the test data. The conversations between female speakers are found to be more easy for the overlapping speech detection task rather than ones between speakers of different genders.

Chapter 7

Conclusions

This thesis has investigated the task of speaker diarization for audio recordings of Broadcast News (BN) and meetings. The speaker diarization is a process of annotating an input audio stream according to speaker identity, thus providing an answer to the question "who spoke when". The work brought out in this thesis follows the assumption that both the number of speakers and the speakers voice are unknown *a priori*. In addition, this thesis has also explored the problem of overlapping speech detection in telephone conversations, although the information of overlapping speech has not yet been exploited by the presented speaker diarization systems.

The implemented speaker diarization system for Broadcast News is based on the partitioning system initially developed at the LIMSI laboratory as a preprocessing stage of the automatic BN transcription system. This baseline system optimizes segment boundaries and cluster labeling jointly via alternating a Viterbi segmentation and an agglomerative clustering iteratively. Since the baseline system is designed to provide a starting point to transcription systems, it is required to give accurate segment boundaries, thus helping to reduce the word errors in automatic transcriptions. However, the speaker diarization task has different objectives compared to the data partitioning for automatic speech transcriptions. Although the speaker diarization also aims to produce homogeneous speech segments, the main objectives are the purity and the correct labeling of the segments. Errors such as having more than one cluster for a given speaker, or conversely, merging the segments of two different speakers into one cluster, are penalized more heavily.

In order to ameliorate speaker diarization performance, some improvements to the baseline system have been implemented during this thesis work. The first modification consists of replacing the iterative segmentation /clustering procedure with an agglomerative BIC clustering based on single full-covariance Gaussian models. The BIC clustering uses a limited parameter distribution model per cluster, since it has to deal with short duration segments at the start of the clustering procedure. The available amount of data in the clusters increases after several merges between clusters, thus allowing to use more complex modeling techniques. The second clustering is then carried out for recombining the clusters resulting from the BIC clustering, relying on techniques

commonly used in speaker identification domain: feature normalization and MAP adaptation from the matching UBM with a large number of Gaussian mixtures.

The improved BN speaker diarization system has been used to participate to the NIST 2004 fall Rich Transcription (RT-04F) evaluation and also the ESTER evaluation. Although the Broadcast News recordings used in both evaluations are in different languages (i.e. RT-04F data in US English and ESTER data in French), the implemented system gave the state-of-the-art diarization performance in the two evaluations. The experiments conducted on the RT-04F development data show that a 70% relative diarization error reduction is achieved by the improved system compared with the baseline system. The results obtained on both the RT-04F and ESTER development data display also that performing the SID clustering reduces significantly the diarization error compared with the system using only the BIC clustering stage. On the test data of both evaluations, the proposed system has similar diarization error approximately of 9% when the system parameters are optimized for the dataset. In the RT-04F evaluation, a further error reduction of 0.6% was achieved by removing inter-word silences based on the word segmentation output by the LIMSI speech-to-text system.

The performance of the improved diarization system is found to depend largely on the setting of system parameters (i.e. the BIC penalty weight λ and the SID clustering threshold δ) that need to be tuned on the specific dataset. A preliminary study on the system robustness is introduced in this thesis, concentrating on the correlation between the optimal values of the system parameters and the durations of BN shows. The examined data was extracted from the ESTER training data and is composed of 20 1-hour BN shows from the source of "France Inter". The contrast experiments were carried out on the subset consisting of the first 30 minutes and the second 30 minutes of shows. The duration of input audio is not found to be a crucial factor that influences greatly the optimal setting of system parameters, even though the optimal SID clustering threshold tends to be smaller on short duration shows. The results obtained using the BIC clustering imply that the BIC penalty term needs to be revised by introducing the factor related to characteristics of audio recordings, besides feature dimensionality and data size.

The improved speaker diarization system for broadcast news is constructed by using separate modules, thus making the system easy to be adapted to new domain via necessary modifications to the blocks. As the meeting domain has received recently more research attention for speech technologies, our speaker diarization system has been adapted from broadcast news to meetings. At the beginning of this adaptation process, the differences between BN and meetings were analyzed in terms of several specific characteristics present in audio. The speech activity detector used in the BN diarization system is based on a Viterbi decoding using Gaussian mixture models representing individually speech, noisy speech, speech over music, pure music, silence and advert. However, the meeting data usually do not contain some types of these acoustic sources such as pure music, silence and advert. Moreover, considering that there are more short speech segments included in meetings than BN shows, a speech activity detection (SAD) algorithm suitable for meetings is thus needed. The proposed SAD method relies on the smoothed log-likelihood ratio (LLR) between the speech and non-speech models that were trained on some lecture data. Different prior probabilities may be given to the speech and non-speech models when performing

the LLR-based speech detector.

In order to compare the performance between the BN diarization system and the developed lecture diarization system, experiments were conducted on the lecture development data used in the NIST Rich Transcription 2006 Spring (RT-06S) meeting recognition evaluation. The LLR-based speech detector provides a significant reduction on the SAD error up to 58% relative to the Viterbi decoding based speech detector. On the RT-06S lecture evaluation data, the diarization system integrating the LLR-based SAD provides an overlap DER of 24.5% on the Single Distant Microphone (SDM) condition, with a similar diarization result obtained on the Multiple Mark III Microphone Arrays (MM3A) condition. A better diarization result of 21.5% DER is obtained on the Multiple Distant Microphone (MDM) condition. The possible reason for this decrease in diarization error is that the randomly selected single microphone signal used for the MDM condition has better quality than that of the SDM condition.

Some further improvements have been implemented in this thesis for developing a common diarization system for both conference and lecture meetings. Since the lecture diarization system provides always high SAD error, some different acoustic representations with various normalization techniques were investigated within the LLR-based speech activity detector. A new energy normalization method taking into account voicing factor was proposed, although the experiments obtained on the RT-07S development data show that the variance normalization technique gives better SAD error than the proposed normalization algorithm. UBMs trained on different types of acoustic features were also examined in the SID clustering stage. Both the speech/non-speech models used in SAD module and UBMs used in SID clustering stage were trained on an union of several previous RT conference datasets, with the corresponding forced alignment transcription references. In addition, the meeting diarization system employed also the beamformed signals generated by the ICSI delay&sum signal enhancement system for the MDM audio input condition.

The adapted meeting speaker diarization system provides a diarization error of 25.6% on the RT-07S lecture SDM evaluation data. Although no improvement on the diarization performance is obtained between the RT-06S and RT-07S lecture evaluation data, it should be noted that the RT-07S lecture data includes more interactions between meeting participants than the RT-06S data, thus bringing more difficulties to the speaker diarization task. A relatively higher diarization error of 29.5% is obtained on the RT-07S conference SDM evaluation data. Using the beamformed signals from all available MDM channels reduces the diarization error to 26.1% on the conference evalaution data. However, for the RT-07S lecture evaluation data, the diarization error rate obtained using the beamformed MDM signals remains very closely to that obtained on the SDM signal.

A preliminary investigation into overlapping speech detection in telephone speech was also carried out during this thesis work. Since each conversation side is recorded in a separate channel, it is possible to generate overlapping speech by directly mixing the speech signals from both speakers. This provides sufficient training data to build a overlapping speech model. Then a log-likelihood ratio based detection algorithm can be performed using the overlapping speech and non-overlapping speech models. Experiments were conducted on conversational telephone

speech, where the first half of each conversation was used as training data and the rest served as test data. Standard cepstral features were compared with autocorrelation features within the LLR-based detection framework. The experimental results show that using autocorrelation features provides better detection performance than cepstral features. Their combination gives similar detection results as that are obtained using autocorrelation features, with the lowest EER of 33.6% for overlapping speech vs. non-overlapping speech decision.

The current researches on the speaker diarization have shown good performance for broadcast news data, while less good diarization results can be obtained on meeting recordings. There is a large room for improving the performance of our meeting diarization system. As multiple channel inputs are normally available for a meeting recording, it is possible to use multi-channel information for helping the speaker diarization. For example, the models trained on inter-channel delays can be combined with acoustic models in the BIC clustering [Pardo *et al.*, 2006a; Pardo *et al.*, 2006b]. Another characteristic of meeting data is to contain a large number of short silence portions. The clustering models can be corrupted by these silence, thus degrading the clustering performance. A purification algorithm aiming to further remove the non-speech frames existent in compared clusters can be used to help the speaker clustering [Anguera *et al.*, 2006b].

Another problem remaining in the speaker diarization domain is the overlapping speech detection. As a very large number of overlapping speech portions is present in meeting recordings, this problem has received more and more attentions. From the NIST RT-06S diarization evaluation, the primary metric has been calculated over the whole audio stream including overlapping portions. Diarization systems are thus required to detect the segment where multiple speakers are talking at the same time and produce the identity corresponding to each speaker. Several researchers have proposed various methods to detect overlapping speech including also the author of this thesis, while no significant gains has been achieved so far in the speaker diarization domain. The speaker location information that can be extracted from the multiple channel inputs may be helpful to overlapping speech detection.

From the experiments presented in this thesis, the parameters existing in diarization systems are found to be sensitive to the processed data and need to be tuned according to development data. The mismatch between development data and test data influences the diarization performance. Improving the robustness of diarization systems is an interest research direction in the future. The potential work on this topic is to make diarization systems selecting optimal values of parameters automatically from the data. To do this, further investigations are needed to find out the underlying correlation between system parameters and acoustic characteristics of audio recordings. Other possible way for improving system robustness is to reduce the number of system parameters or training data used to build acoustic models.

Appendix A

Likelihood ratio for Gaussian models

Let $X = \{x_1, \cdots, x_N\}$ be a sequence of d-dimensional feature vectors, assuming that X is generated by a multivariate Gaussian process $\mathcal{N}(\mu, \Sigma)$, the probability density is parameterized by $d \times 1$ mean vector μ and $d \times d$ covariance matrix Σ as:

$$p(x; \mu, \Sigma) = \frac{1}{2\pi^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\Big) \tag{A.1}$$

The likelihood of the Gaussian model $\mathcal{N}(\mu, \Sigma)$ for the observed sequence X is:

$$L(X; \mu, \Sigma) = \prod_{i=1}^{N} \frac{1}{2\pi^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\Big) \tag{A.2}$$

The logarithm of this likelihood is given as:

$$\log L(X; \mu, \Sigma) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)) \tag{A.3}$$

A basic hypothesis test is considered as:

- H_0: the vector sequence X is modeled by a single multivariate Gaussian

  $$X = \{x_1, \cdots, x_N\} \sim \mathcal{N}(\mu, \Sigma)$$

- H_1: the vector sequence X is modeled by two different multivariate Gaussians

  $$X_1 = \{x_1, \cdots, x_{N_1}\} \sim \mathcal{N}(\mu_1, \Sigma_1) \text{ and } X_2 = \{x_{N_1+1}, \cdots, x_N\} \sim \mathcal{N}(\mu_2, \Sigma_2)$$

  with $N_2 = N - N_1$.

The likelihood ratio of this hypothesis test is given by:

$$R = \frac{L(X;\mu,\Sigma)}{L(X_1;\mu_1,\Sigma_1)L(X_2;\mu_2,\Sigma_2)} \tag{A.4}$$

The logarithm of this statistic is used giving the log-likelihood ratio

$$\begin{aligned}
\log R &= \log L(X;\mu,\Sigma) - \log L(X_1;\mu_1,\Sigma_1) - \log L(X_2;\mu_2,\Sigma_2) \\
&= -\frac{N}{2}\log|\Sigma| + \frac{N_1}{2}\log|\Sigma_1| + \frac{N_2}{2}\log|\Sigma_2| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \\
&\quad + \frac{1}{2}\sum_{i=1}^{N_1}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1) + \frac{1}{2}\sum_{i=1}^{N_2}(x_i-\mu_2)^T\Sigma_2^{-1}(x_i-\mu_2)
\end{aligned} \tag{A.5}$$

Let $\sigma[m,n]$ represent the entry of the matrix Σ^{-1} that lies in the m'th row and the n'th column, and let $x_i[k]$ and $\mu[k]$ be the k'th component of the i'th vector and the mean vector μ respectively, then we can write

$$\begin{aligned}
\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) &= \sum_{i=1}^{N}\Big(\sum_{n=1}^{d}\sum_{m=1}^{d}(x_i[m]-\mu[m])\sigma[m,n](x_i[n]-\mu[n])\Big) \\
&= \sum_{n=1}^{d}\sum_{m=1}^{d}\sigma[m,n]\Big(\sum_{i=1}^{N}(x_i[m]-\mu[m])(x_i[n]-\mu[n])\Big)
\end{aligned} \tag{A.6}$$

Assuming the covariance matrix S is a maximum liklihood estimation from the sample X, then this covariance matrix is computed as:

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T \tag{A.7}$$

with the element $s[m,n]$

$$s[m,n] = \frac{1}{N}\sum_{i=1}^{N}(x_i[m]-\mu[m])(x_i[n]-\mu[n]) \tag{A.8}$$

Thus (A.6) can be rearranged as:

$$\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = N\sum_{n=1}^{d}\sum_{m=1}^{d}\sigma[m,n]s[m,n] \tag{A.9}$$

Since the matrix Σ^{-1} is a symmetric matrix, (A.9) can be reformulated

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) &= N\sum_{n=1}^{d}\sum_{m=1}^{d}\sigma[n,m]s[m,n] \\ &= Ntr(\Sigma^{-1}S) \end{aligned} \tag{A.10}$$

where tr is the trace of the matrix.

We assume that the estimation of the maximum likelihood from the sample X is perfect, i.e. $S = \Sigma$, thus

$$\Sigma^{-1}S = I_d \tag{A.11}$$

where I_d is the identity matrix.

With this assumption, we simplify (A.10)

$$\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = Nd \tag{A.12}$$

Using (A.12), the log-likelihood ratio (A.5) can be estimated as:

$$\begin{aligned}\log R &= -\frac{N}{2}\log|\Sigma| + \frac{N_1}{2}\log|\Sigma_1| + \frac{N_2}{2}\log|\Sigma_2| - \frac{N}{2}d + \frac{N_1}{2}d + \frac{N_2}{2}d \\ &= -\frac{N}{2}\log|\Sigma| + \frac{N_1}{2}\log|\Sigma_1| + \frac{N_2}{2}\log|\Sigma_2| \end{aligned} \tag{A.13}$$

Appendix B

List of personal publications

[1] C. Barras, X. Zhu, S. Meignier and J.-L. Gauvain, "Improving Speaker Diarization", In *Proceedings of FAll 2004 Rich Transcription Workshop (RT-04)*, Palisades, New York, USA, November 2004

[2] X. Zhu, C. C. Leung, C. Barras, L. Lamel and J.-L. Gauvain, "Speech Activity Detection and Speaker Identification for CHIL", In *Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'05)*, Edinburgh, July 2005

[3] X. Zhu, C. Barras, S. Meignier and J-L. Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization", In *Proceedings of the 9th European Conference on Speech Communication and Technology (ISCA Interspeech'05)*, pp. 2441-2444, Lisboa, September 2005

[4] C. Barras, X. Zhu, S. Meignier and J-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast New", In the *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1505-1512, September 2006

[5] C. Barras, X. Zhu, J.-L. Gauvain and L. Lamel, "The CLEAR'06 LIMSI Aoustic Speaker Identification System for CHIL Seminars", In *Proceedings of workshop on Classification of Events, Activities and Relationships (CLEAR'06)*, Southampton, UK, April 2006

[6] X. Zhu, C. Barras, L. Lamel and J.-L. Gauvain, "Speaker Diarization: from Broadcast News to Lectures", In *Proceedings of 3nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'06)*, Washington DC, USA, May 2006

[7] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, X. Zhu, "The LIMSI 2006 TC-STAR Transcription Systems", In *Proceedings of 2007 International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP'03)*, Honolulu,

Hawaii, USA, April 2007

[8] X. Zhu, C. Barras, L. Lamel, J.-L. Gauvain, "Multi-Stage Speaker Diarization for Conference and Lecture Meetings", In *Rich Transcription 2007 Meeting Recognition Evaluation Workshop (RT-07)*, Baltimore, USA, May 2007

[9] C. Barras, X. Zhu, C. C. Leung, J.-L. Gauvain, L. Lamel, "Acoustic Speaker Identification: The LIMSI CLEAR'07 System", In *Proceedings of workshop on Classification of Events, Activities and Relationships (CLEAR'07)*, Baltimore, USA, May 2007

Bibliography

[Adami *et al.*, 2002] A. G. Adami, S. S. Kajarekar, and H. Hermansky. A new speaker change detection method for two-speaker segmentation. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2002)*, volume 4, pages 3908–3911, Orlando, FL, USA, May 2002.

[Adda *et al.*, 2007] G. Adda, M. Adda-Decker, C. Barras, P. Boula de Mareüil, B. Habert, and P Paroubek. Speech overlap and interplay with disfluencies in political interviews. In *ParaLing'07: International workshop on Paralinguistic speech - between models and data*, Saarbrücken, August 2007.

[Ajmera and Wooters, 2003] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Automatic Speech Recognition and Understanding (IEEE, ASRU 2003)*, pages 411–416, St. Thomas, U.S. Virgin Islands, November 2003.

[Ajmera *et al.*, 2002] J. Ajmera, I. A. McCowan, and H. Bourlard. BIC revisited for speaker change detection. Technical report, IDIAP Research Report, October 2002.

[Akaike, 1974] H. Akaike. A new look at the statistical identification model. *In IEEE Transactions Automatic Control*, 19:716–723, 1974.

[Anguera *et al.*, 2005a] X. Anguera, C. Wooters, and J. Hernando. Speaker diarization for multiparty meetings using acoustic fusion. In *Automatic Speech Recognition and Understanding (IEEE, ASRU 2005)*, San Juan, Puerto Rico, November 2005.

[Anguera *et al.*, 2005b] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo. Robust speaker segmentation for meeting: The ICSI-SRI spring 2005 diarization system. In *Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2005)*, Edinburgh, UK, July 2005.

[Anguera *et al.*, 2006a] X. Anguera, M. Aguilo, C Wooters, C. Nadeu, and J. Hernando. Hybrid speech/non-speech detector applied to speaker diarization of meetings. In *2006: A Speaker Odyssey. The Speaker Recognition Workshop (ISCA, Odyssey 2006)*, Puerto Rico, USA, June 2006.

[Anguera *et al.*, 2006b] X. Anguera, C. Wooters, and J. Hernando. Purity algorithm for speaker diarization of meetings data. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2006)*, Toulouse, France, May 2006.

[Anguera *et al.*, 2006c] X. Anguera, C. Wooters, and J. M. Pardo. Robust speaker diarization for meetings: ICSI RT06s evaluation system. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2006)*, pages 1674–1677, Pittsburgh, USA, September 2006.

[Anguera, 2005] X. Anguera. XBIC: real-time cross probabilities measure for speaker segmentation. Technical report, ICSI-Berkeley Technique Report, August 2005.

[Anguera, 2006] X. Anguera. *Robust Speaker Diarization for Meetings*. PhD thesis, Univeritat Politecnica de Catalunya, 2006.

[Atal and Hanauer, 1971] B. Atal and S. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *In Journal of the Acoustical Society of America*, 50(2):637–655, 1971.

[Backer, 1975] J. K. Backer. The DRAGON system–an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):24–29, 1975.

[Barras and Gauvain, 2003] C. Barras and J.-L. Gauvain. Feature and score normalization for speaker verification of cellular data. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2003)*, volume 2, pages 49–52, Hong Kong, China, April 2003.

[Barras *et al.*, 2004] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Improving speaker diarization. In *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, November 2004.

[Barras *et al.*, 2006] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multi-stage speaker diarization of broadcast new. *IEEE Transactions on audio, Speech and Language Processing*, 14(5):1505–1512, September 2006.

[Baum and Petrie, 1966] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[Ben *et al.*, 2004] M. Ben, M. Betser, F. Bimbot, and G. Gravier. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adaptaed GMMs. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2004)*, Jeju, Korea, October 2004.

[Bilmes, 1998] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical report, International Computer Science Institute, April 1998.

[Bonastre *et al.*, 2000] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, and C. J. Wellekens. A speaker tracking system based on speaker turn detection for NIST evaluations. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2000)*, Istanbul, Turkey, November 2000.

[Bregman, 1994] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1994.

[Campbell, 1997] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[Canseco-Rodriguez *et al.*, 2004] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain. Speaker diarization from speech transcripts. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2004)*, pages 1272–1275, Jeju, Korea, October 2004.

[Canseco-Rodriguez *et al.*, 2005] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain. A comparative study using manual and automatic transcription for diarization. In *Automatic Speech Recognition and Understanding (IEEE, ASRU 2005)*, San Juan, Puerto Rico, November 2005.

[Canseco-Rodriguez, 2006] L. Canseco-Rodriguez. *Speaker Diarization in Broadcast News*. PhD thesis, LIMSI, Universite Paris Sud, 2006.

[Cettolo *et al.*, 2005] M. Cettolo, M. Vescovi, and R. Rizzi. Evaluation of BIC-based algorithms for audio segmentation. *Computer Speech and Language*, 19:147–170, 2005.

[Cettolo, 2000] M. Cettolo. Segmentation, classification and clustering of an italian broadcast news corpus. In *Conference on Content-Based Multimedia Information Access (RIAO 2000)*, Paris, France, April 2000.

[Chen and Gopalakrishnan, 1998a] S. Chen and P. Gopalakrishnan. Clustering via the Bayesian information criterion with applications in speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 1998)*, Seattle, Washington, USA, May 1998.

[Chen and Gopalakrishnan, 1998b] S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, February 1998.

[Cooke and Ellis, 2001] M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35(3-4):141–177, October 2001.

[Davis and Mermelstein, 1980] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[De Cheveigné, 2006] A. De Cheveigné. Multiple F0 estimation. In DeLiang Wang and Guy J. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pages 65–70. Wiley/IEEE Press, 2006.

[Delacourt and Wellekens, 1999] P. Delacourt and C. Wellekens. Audio data indexing: use of second-order statistics for speaker-based segmentation. In *IEEE International Conference on Nultimedia, computing and Systems*, 1999.

[Delacourt and Wellekens, 2000] P. Delacourt and C. Wellekens. DISTBIC: a speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1-2):111–126, September 2000.

[Delacourt *et al.*, 1999a] P. Delacourt, D. Kryze, and C. Wellekens. Detection of speaker changes in an audio document. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 1999)*, September 1999.

[Delacourt *et al.*, 1999b] P. Delacourt, D. Kryze, and C. Wellekens. Speaker-based segmentation for audio data indexing. In *ESCA Workshop: Accessing Information in Audio Data*, 1999.

[Delacourt, 2000] P. Delacourt. *La Segmentation et le Regroupement par Locuteurs pour l'Indexation de Document Audio*. PhD thesis, ENST-Eurecom, 2000.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Acoustical Society of America*, 39:1–38, 1977.

[Doddington *et al.*, 2000] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, 31:225–254, 2000.

[El Khoury *et al.*, 2007] E. El Khoury, C. SENAC, and R. ANDRE OBRECHT. Speaker diarization: towards a more robust and portable system. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2007)*, volume 4, pages 489–492, Honolulu, Hawaii, USA, April 2007.

[Fiscus *et al.*, 2006] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo. The Rich Transcription 2006 spring meeting recognition evaluation. In *Proceedings of 3nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006)*, Washington DC, USA, May 2006.

[Fredouille and Senay, 2006] C. Fredouille and G. Senay. Technical improvements of the E-HMM based speaker diarization system for meeting records. In *Proceedings of 3nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006)*, Washington DC, USA, May 2006.

[Furui, 1981a] S. Furui. Cepstral analysis technique for speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, Avril 1981.

[Furui, 1981b] S. Furui. Comparison of speaker recognition methods using static features and dynamic faetures. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):342–350, 1981.

[Furui, 1986] S. Furui. Speaker-independent isolated word recognition using dynamic features of the speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, February 1986.

[Gallardo-Antolin *et al.*, 2006] A. Gallardo-Antolin, X. Anguera, and C. Wooters. Multi-stream speaker diarization system for the meetings domain. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2006)*, pages 2186–2189, Pittsburgh, USA, September 2006.

[Galliano *et al.*, 2005] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of 9th European Conference on Speech Communication and Technology (ISCA, Interspeech 2005)*, pages 1149–1152, Lisbon, Portugal, Spetember 2005.

[Gangadharaiah *et al.*, 2004] R. Gangadharaiah, B. Narayanaswamy, and N. Balakrishnan. A novel method for two-speaker segmentation. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2004)*, Jeju, Korea, October 2004.

[Gauvain and Lee, 1994] J.-L. Gauvain and C. H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 22:291–298, April 1994.

[Gauvain *et al.*, 1998] J.-L. Gauvain, L. Lamel, and G. Adda. Partitioning and transcription of broadcast news data. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 1998)*, pages 1335–1338, December 1998.

[Gauvain *et al.*, 2001] J.-L. Gauvain, L. Lamel, and G. Adda. Audio partitioning and transcription for broadcast data indexation. *Multimedia Tools and Applications*, 14:187–200, 2001.

[Gauvain *et al.*, 2002] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[Gauvain *et al.*, 2003] J.-L. Gauvain, L. Lamel, G. Adda, L. Z. Chen, and F. Lefevre. Conversational telephone speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2003)*, pages 212–215, Hong Kong, April 2003.

[Gish *et al.*, 1991] H. Gish, M.-H. Siu, and R. Rohlicek. Segregation of speaker for speech recognition and speaker identification. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 1991)*, volume 2, pages 873–876, Toronto, Canada, May 1991.

[Graff *et al.*, 2001] David Graff, Kevin Walker, and David Miller. Switchboard cellular part I transcribed audio. Linguistic Data Consortium, 2001.

[Gravier *et al.*, 2004] G. Gravier, J.-F. Bonastre, S. Galliano, K. Geoffrois, E. amd Mc Tait, and K. Choukri. The ESTER evaluation campaign of rich transcription of French broadcast news. In *Language Evaluation and Resources Conference (LREC 2004)*, Lisbon, Portugal, May 2004.

[Hain and Woodland, 1998] T. Hain and P. C. Woodland. Segmentation and classification of broadcast news audio. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, November 1998.

[Hain *et al.*, 1998] 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. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 133–137, Landsdowne, VA, USA, February 1998.

[Harris *et al.*, 1999] M. Harris, X. Aubert, R. Haeb-Umbach, and P. Beyerlein. A study of broadcast news audio stream segmentation and segment clustering. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 1999)*, pages 1027–1030, Budapest, Hungary, September 1999.

[Harris, 1978] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[Hermanksy and Morgan, 1994] H. Hermanksy and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

[Hermansky, 1990] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *In Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[Huang *et al.*, 1990] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, Edinburgh, UK, 1990.

[Hung *et al.*, 2000] J. W. Hung, H. M. Wang, and L. S. Lee. Automatic metric-based speech segmentation for broadcast news via principal component analysis. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2000)*, Beijing, China, 2000.

[Istrate *et al.*, 2005a] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J.-F. Bonastre. NIST RT05S evaluation: pre-processing techniques and speaker diarization on multiple microphone meetings. In *Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2005)*, Edinburgh, UK, July 2005.

[Istrate *et al.*, 2005b] D. Istrate, N. Scheffer, C. Fredouille, and J.-F. Bonastre. Broadcast news speaker tracking for ESTER 2005 campaign. In *Proceedings of 9th European Conference on Speech Communication and Technology (ISCA, Interspeech 2005)*, pages 2445–2448, Lisbon, Portugal, Spetember 2005.

[Itakura and Saito, 1968] F. Itakura and S. Saito. Analysis synthesis telephony based upon the maximum likelihood method. In *6th International Congress on Aoustics*, pages C–5–5, Tokyo, 1968.

[Jelinek, 1976] F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 4(64):532–556, 1976.

[Jin *et al.*, 1997] H. Jin, F. Kubala, and R. Schwartz. Automatic speaker clustering. In *Proceedings of the DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.

[Jin *et al.*, 2004] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel. Speaker segmentation and clustering in meetings. In *Proceedings of Spring 2004 Rich Transcription Workshop (RT-04S)*, Montreal, Canada, May 2004.

[Johnson and Woodland, 1998] S. E. Johnson and P. C. Woodland. Speaker clustering using direct maximisation of the MLLR-adapted likelihood. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 1998)*, volume 5, pages 1775–1779, Sydney, Australia, November 1998.

[Johnson, 1999] S. E. Johnson. Who spoke when? - automatic segmentation and clustering for determining speaker turns. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 1999)*, Budapest, Hungary, September 1999.

[Junqua and Haton, 1996] J. C. Junqua and J. P. Haton. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer, 1996.

[Kemp *et al.*, 2000] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2000)*, pages 1423–1426, Istanbul, Turkey, November 2000.

[Koolwaaij and Boves, 2000] J. Koolwaaij and L. Boves. Local normalization and delayed decision making in speaker detection and tracking. *Digital Signal Processing*, 10(1/2/3):113–132, 2000.

[Kubala *et al.*, 1997] F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, and J. Makhoul. BBN Byblos Hub-4 transcription system. In *Proceedings of the Speech Recognition Workshop*, pages 90–93, 1997.

[Lamel *et al.*, 1981] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4):777–785, 1981.

[Laskowski and Schultz, 2006] K. Laskowski and T. Schultz. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2006)*, pages 993–996, Toulouse, France, May 2006.

[Leggetter and Woodland, 1995] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[Levinson, 1983] S. C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

[Liu and Kubala, 1999] D. Liu and F. Kubala. Fast speaker change detection for broadcast news transcription and indexing. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 1999)*, volume 3, pages 1031–1034, Budapest, Hungary, September 1999.

[Liu and Kubala, 2003] D. Liu and F. Kubala. Online speaker clustering. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2003)*, volume 1, pages 572–575, Hongkong, China, April 2003.

[Lu and Zhang, 2002] L. Lu and H. J. Zhang. Real-time unsupervised speaker change detection. In *Proceedings of International Conference on Pattern Recognition (ICPR 2002)*, volume 2, Quebec City, Canada, August 2002.

[Lu *et al.*, 2001] L. Lu, H. Jiang, and H. J. Zhang. A robust audio classification and segmentation method. In *Proceedings of the 9th ACM International Conference on Multimedia (ACM 2001)*, pages 203–211, Ottawa, Ontario, Canada, 2001.

[Makhoul, 1973] J. Makhoul. Spectral analysis of speech by linear prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 21(3):140–148, 1973.

[Makhoul, 1975] J. Makhoul. Linear prediction: a tutorial review. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 63(4):561–580, 1975.

[Malegaonkar *et al.*, 2006] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna. Unsupervised speaker change detection using probabilistic pattern matching. *IEEE Signal Processing Letters*, 13(8):509–512, August 2006.

[Mariani, 2002] J. Mariani. *Analyse, Synthese et Codage de la Parole*. Hermes Science, Paris, France, 2002.

[Markel and Gray, 1976] J. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer-Verlag, New York, NY, USA, 1976.

[McLachlan, 1988] G. McLachlan. *Mixture Models*. Marcel Dekker, New York, NY, USA, 1988.

[Meignier *et al.*, 2000] S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin. Evolutive HMM for speaker tracking system. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2000)*, pages 1177–1180, Istanbul, Turkey, November 2000.

[Meignier *et al.*, 2001] S. Meignier, J.-F. Bonastre, and S. Igounet. E-HMM approach for learning and adapting sound models for speaker indexing. In *2001: A Speaker Odyssey. The Speaker Recognition Workshop (ISCA, Odyssey 2001)*, pages 175–180, Chania, Crete, Greece, June 2001.

[Meignier *et al.*, 2006] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech and Language*, 20(2-3):303–330, 2006.

[Meignier, 2002] S. Meignier. *Indexation en Locuteurs de Documents Sonores: Segmentation d'un Document et Appariement d'une Collection*. PhD thesis, Universite d'Avignon et des Pays de Vaucluse, 2002.

[Mirghafori and Wooters, 2006] N. Mirghafori and C. Wooters. Nuts and flakes: A study of data characteristics in speaker diarization. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2006)*, pages 1017–1020, Toulouse, France, May 2006.

[Moh *et al.*, 2003] Y. Moh, P. Nguyen, and J.-C. Junqua. Towards domain independent speaker clustering. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2003)*, Hong Kong, China, April 2003.

[Moraru *et al.*, 2004a] D. Moraru, L. Besacier, and E. Castelli. Using a priori information for speaker diarization. In *2004: A Speaker Odyssey. The Speaker Recognition Workshop (ISCA, Odyssey 2004)*, pages 355–362, Toledo, Spain, May 2004.

[Moraru *et al.*, 2004b] D. Moraru, L. Besacier, S. Meignier, C. Fredouille, and J.-F. Bonastre. Speaker diarization in the ELISA consortium over the last 4 years. In *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, November 2004.

[Moraru *et al.*, 2004c] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 Rich Transcription evaluation. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2004)*, Montreal, Canada, May 2004.

[Moraru *et al.*, 2005] D. Moraru, M. Ben, and G. Gravier. Experiments on speaker tracking and segmentation in radio broadcast news. In *Proceedings of 9th European Conference on Speech Communication and Technology (ISCA, Interspeech 2005)*, pages 3049–3052, Lisbon, Portugal, Spetember 2005.

[Moraru, 2004] D. Moraru. *Segmentation en Locuteur de Documents Audio et Audiovisuels: Application a la Recherche d'Information Multimedia*. PhD thesis, Institut National Polytechnique de Grenoble, 2004.

[Mori and Nakagawa, 2001] K. Mori and S. Nakagawa. Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition. In *Proceedings of*

International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2001), volume I, pages 413–416, Salt Lake City, Utah, USA, 2001.

[Nakagawa and Suzuki, 1993] S. Nakagawa and H. Suzuki. A new speech recognition method based on VQ-distortion measure and HMM. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 1993)*, pages 676–679, 1993.

[Nguyen *et al.*, 2004] L. Nguyen, S. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J.-L. Gauvain, G. Adda, H. Schwenk, and F. Lefevre. The 2004 BBN/LIMSI 10xRT english broadcast news transcription system. In *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, Nov 2004.

[Nishida and Ariki, 1998] M. Nishida and Y. Ariki. Real time speaker indexing based on subspace method: applications to TV news articles and debate. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 1998)*, pages 1347–1350, December 1998.

[Nishida and Kawahara, 2003] M. Nishida and T Kawahara. Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2003)*, HongKong, April 2003.

[NIST, 2000] NIST. The 2000 NIST Speaker Recognition Evaluation Plan. http://www.nist.gov/speech/tests/spk/2000/doc/spk-2000-plan-v1.0.htm, January 2000.

[NIST, 2003] NIST. The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan. http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf, February 2003. (Version 4, Updated 02/25/2003).

[NIST, 2004] NIST. Fall 2004 Rich Transcription (RT-04F) Evaluation Plan. http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf, August 2004.

[NIST, 2005] NIST. Spring 2005 Rich Transcription (RT-05S) Meeting Recognition Evaluation Plan. http://www.nist.gov/speech/tests/rt/rt2005/spring/rt05s-meeting-eval-plan-v1.pdf, February 2005.

[NIST, 2006] NIST. Spring 2006 Rich Transcription (RT-06S) Meeting Recognition Evaluation Plan. http://www.nist.gov/speech/tests/rt/rt2006/spring/docs/rt06s-meeting-eval-plan-V2.pdf, February 2006.

[NIST, 2007] NIST. Spring 2007 Rich Transcription (RT-07S) Meeting Recognition Evaluation Plan. http://www.nist.gov/speech/tests/rt/rt2007/spring/docs/rt07s-meeting-eval-plan-V2.pdf, February 2007.

[NISTRT,] NISTRT. Benchmark Tests: Rich Transcription Evaluations. http://www.nist.gov/speech/tests/rt/index.htm.

[NISTSRE,] NISTSRE. Benchmark Tests: Speaker Recognition Evaluations. http://www.nist.gov/speech/tests/spk/index.htm.

[Oppenheim and Schafer, 1968] A. Oppenheim and R. Schafer. Homomorphic analysis of speech. *In IEEE Trasactions on Audio and Electroacoustics*, 16(2):221–226, 1968.

[Oppenheim and Schafer, 1975] A. Oppenheim and R. Schafer. *Digital Signal Processing*. Prentice Hall, 1975.

[O'Shaughnessy, 1987] D. O'Shaughnessy. *Speech Communication: Humain and Machine*. Addison Wesley, New York, NY, USA, 1987.

[Pardo *et al.*, 2006a] J. M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multi-microphone meetings: using only between-channel differences. In *Proceedings of 3nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006)*, Washington DC, USA, May 2006.

[Pardo *et al.*, 2006b] J. M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multiple distant microphones meetings: mixing acoustic features and inter-channel time differences. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2006)*, pages 2194–2197, Pittsburgh, USA, September 2006.

[Pelecanos and Sridharan, 2001] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *2001: A Speaker Odyssey. The Speaker Recognition Workshop (ISCA, Odyssey 2001)*, June 2001.

[Pfau *et al.*, 2001] T. Pfau, D. P. W. Ellis, and A. Stolcke. Multispeaker speech activity detection for the ICSI meeting recorder. In *Automatic Speech Recognition and Understanding (IEEE, ASRU 2001)*, Trento, Italy, December 2001.

[Picone, 1993] J. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, 1993.

[Rabiner and Juang, 1993] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[Rabiner and Schafer, 1978] L. R. Rabiner and R. Schafer. *Digital Processing of Speech Signal*. Prentice Hall, 1978.

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

[Reynolds and Torres-Carrasquillo, 2004] D. A. Reynolds and Torres-Carrasquillo. The MIT Lincoln laboratory RT-04F diarization systems: applications to broadcast news and telephone conversations. In *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, November 2004.

[Reynolds and Torres-Carrasquillo, 2005] D. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2005)*, Philadelphia, March 2005.

[Reynolds *et al.*, 1998] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. J. McLaughlin, and M. A. Zissman. Blind clustering of speech utterances based on speaker and language characteristics. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, November 1998.

[Reynolds *et al.*, 2000] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3):19–41, 2000.

[Reynolds, 1992] D. A. Reynolds. *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*. PhD thesis, Georgia Institute of Technology, 1992.

[Roch and Cheng, 2004] M. Roch and Y. L. Cheng. Speaker segmentation using the map-adapted Bayesian information criterion. In *2004: A Speaker Odyssey. The Speaker Recognition Workshop (ISCA, Odyssey 2004)*, Toledo, Spain, May 2004.

[Rudasi and Zahorian, 1991] L. Rudasi and S. A. Zahorian. Text-independent talker identification with neural networks. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 1991)*, volume 1, pages 389–392, Toronto, Canada, May 1991.

[Sacks *et al.*, 1974] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, Decembre 1974.

[Schmidt, 1996] M. Schmidt. Identifying speaker with support vector networks. In *Interface 1996 Proceedings*, Sydney, Australia, 1996.

[Schroeder and Campbell, 2000] J. Schroeder and J. Campbell, editors. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*. Academic Press, 2000.

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

[Shriberg *et al.*, 2001] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 2001)*, pages 1359–1362, Aalborg, September 2001.

[Siegler *et al.*, 1997] M. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation and clustering of broadcast news audio. In *Proceedings of the DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.

[Sinha *et al.*, 2005] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland. The Cambridge University March 2005 speaker diarization system. In *Proceedings of 9th European Conference on Speech Communication and Technology (ISCA, Interspeech 2005)*, pages 2437–2440, Lisbon, Portugal, Spetember 2005.

[Solomonoff *et al.*, 1998] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish. Clustering speakers by their voices. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 1998)*, Seattle, Washington, USA, May 1998.

[Soong *et al.*, 1985] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang. A vector quantization approach to speaker recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 1985)*, pages 387–390, Tampa, FL, USA, March 1985.

[Stolcke *et al.*, 2005] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng. The ICSI-SRI spring 2005 speech-to-text evaluation system. In *Proceedings of Rich Transcription 2005 Spring Meeting Recognition Workshop (RT-05)*, Edinburgh, Great Britain, July 2005.

[Ten Bosch *et al.*, 2005] L. Ten Bosch, N. Oostdijk, and L. Boves. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86, September-October 2005.

[Tranter and Reynolds, 2004] S. E. Tranter and D. A. Reynolds. Speaker diarisation for broadcast news. In *2004: A Speaker Odyssey. The Speaker Recognition Workshop (ISCA, Odyssey 2004)*, Toledo, Spain, May 2004.

[Tranter and Reynolds, 2006] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, Speech and Language Processing*, 14(5):1557–1565, September 2006.

[Tranter and Yu, 2003] S. E. Tranter and K. Yu. Diarization for RT-03s at Cambridge University. In *Proceedings of Spring 2003 Rich Transcription Workshop (RT-03)*, Boston, USA, Mai 2003.

[Tranter *et al.*, 2004] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, and P. C. Woodland. The development of the Cambridge University RT-04 diarization system. In *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, November 2004.

[Tritschler and Gopinath, 1999] A. Tritschler and R. Gopinath. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 1999)*, pages 679–682, Budapest, Hungary, Spetember 1999.

[Tsai *et al.*, 2004] W. H. Tsai, S. S. Cheng, and H. M. Wang. Speaker clustering of speech utterances using a voice characteristic reference space. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2004)*, Jeju, Korea, October 2004.

[Tsai *et al.*, 2005] W. H. Tsai, S. S. Cheng, Y. H. Chao, and H. M. Wang. Clustering speech utterances by speaker using eigenvoice-motivated vector space models. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2005)*, volume 1, pages 725–728, Philadelphia, March 2005.

[Van Leeuwen and Konecny, 2007] D. A. Van Leeuwen and M. Konecny. Progress in the AMIDA speaker diarization system for meeting data. In *Proceedings of Rich Transcription 2007 Meeting Recognition Workshop (RT-07)*, Baltimore, USA, May 2007.

[Vandecatseye and Martens, 2003] A. Vandecatseye and J.-P. Martens. A fast, accurate and stream-based speaker segmentation and clustering algorithm. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 2003)*, pages 941–944, Geneva, Swizerland, September 2003.

[Viterbi, 1967] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 2(IT-13):260–269, April 1967.

[Vogt *et al.*, 2003] R. Vogt, J. Pelecanos, and S. Sridharan. Dependence of GMM adaptation on feature post-processing for speaker recognition. In *Proceedings of European Conference on Speech Communication and Technology (ISCA, Eurospeech 2003)*, pages 3013–3016, Geneva, Swizerland, September 2003.

[Wang and Brown, 2006] D. L. Wang and J. Brown, G, editors. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley/IEEE Press, 2006.

[Woodland *et al.*, 1997] P. C. Woodland, M. Gales, D. Pye, and S. Young. The development of the 1996 HTK broadcast news transcription system. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 73–78, Chantilly, VA, USA, February 1997.

[Wooters and Huijbregts, 2007] C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Proceedings of Rich Transcription 2007 Meeting Recognition Workshop (RT-07)*, Baltimore, USA, May 2007.

[Wooters *et al.*, 2004] C. Wooters, J. Fung, B. Peskin, and X. Anguera. Towards robust speaker segmentation: the ICSI-SRI Fall 2004 diarization system. In *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, November 2004.

[Wrigley *et al.*, 2005] S. J. Wrigley, G. J. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multi-channel audio. *IEEE Transactions on Speech and Audio Processing*, 13:84–91, 2005.

[Wu *et al.*, 2003a] T. L. Wu, L. Lu, K. Chen, and Zhang H. J. UBM-based real-time speaker segmentation for broadcasting news. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2003)*, HongKong, April 2003.

[Wu *et al.*, 2003b] T. L. Wu, L. Lu, K. Chen, and Zhang H. J. Universal background models for real-time speaker change detection. In *International Conference on Multimedia Modeling*, 2003.

[Xiang *et al.*, 2002] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath. Short-time gaussianization for robust speaker verification. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2002)*, volume 1, pages 681–684, Orlando, FL, USA, May 2002.

[Young, 1996] S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Transactions on Signal Processing*, 13(55):45–57, 1996.

[Zheng and Yuan, 1988] Y. C. Zheng and B. Z. Yuan. Text-dependent speaker identification using circular hidden Markov model. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 1988)*, volume 1, pages 580–582, New York, NY, USA, April 1988.

[Zhou and Hansen, 2000] B. Zhou and J. H. L. Hansen. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In *Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2000)*, volume 3, pages 714–717, Beijing, China, 2000.

[Zhu *et al.*, 2005] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain. Combing speaker identification and BIC for speaker diarization. In *Proceedings of 9th European Conference on Speech Communication and Technology (ISCA, Interspeech 2005)*, pages 2441–2444, Lisbon, Portugal, Spetember 2005.

[Zhu *et al.*, 2006] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain. Speaker diarization: from broadcast news to lectures. In *Proceedings of 3nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006)*, Washington DC, USA, May 2006.

[Çetin and Shriberg, 2006] Ö. Çetin and E. Shriberg. Errors in meetings: Effects before, during, and after the overlap. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2006)*, Toulouse, France, May 2006.