Multistage Speaker Diarization of Broadcast News

Author :Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain

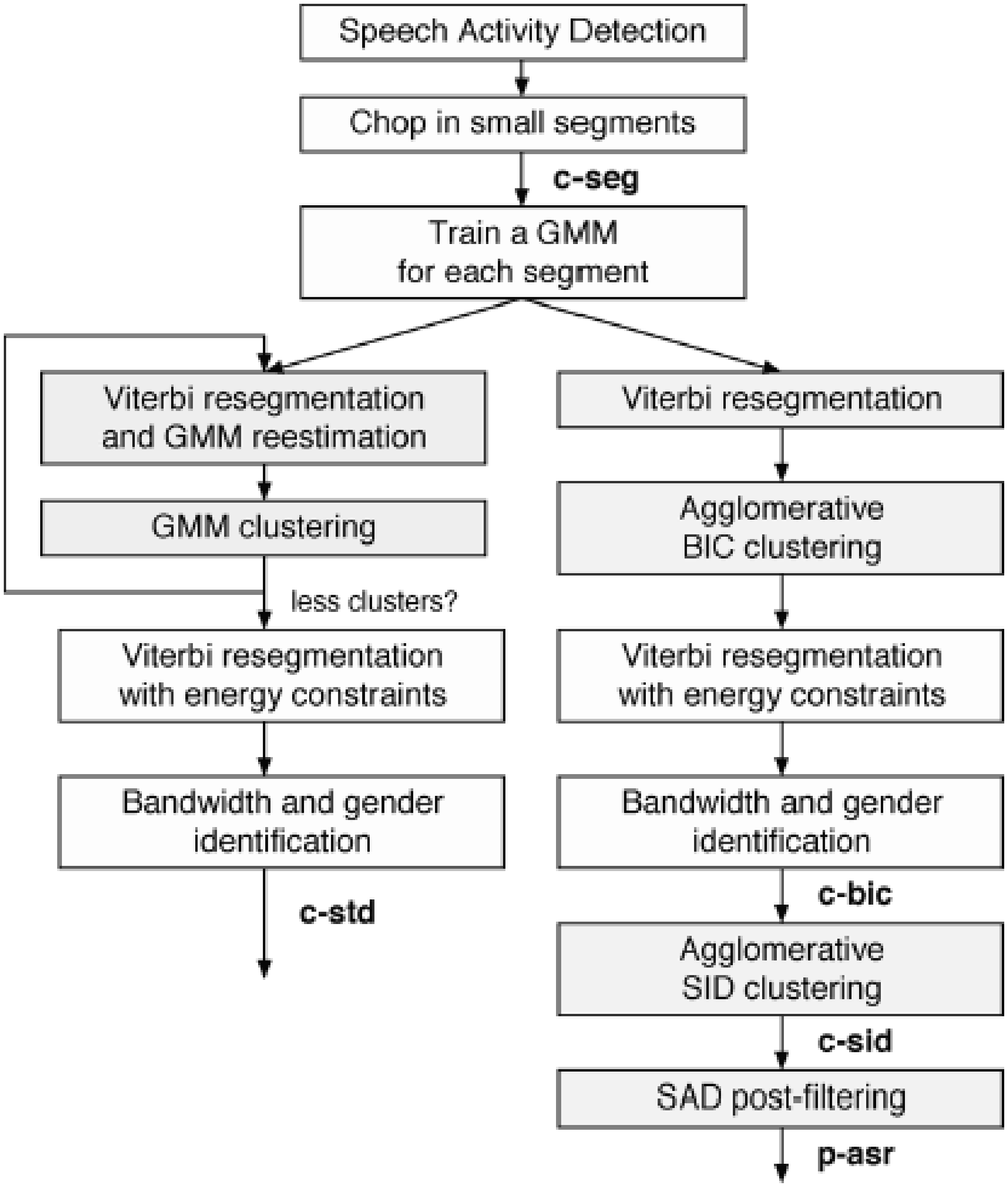
Professor : 陳嘉平

Reporter : 楊治鏞

Introduction

- Speaker diarization, also called speaker segmentation and clustering, is the process of partitioning an input audio stream into homogeneous segments according to speaker identity.

- Audio diarization is a useful preprocessing step for an automatic speech transcription system.


Speech Activity Detection
Chop in small segments
c-seg
Train a GMM for each segment
Viterbi resegmentation and GMM reestimation
GMM clustering
less clusters?
Viterbi resegmentation with energy constraints
Bandwidth and gender identification
c-std
Viterbi resegmentation
Agglomerative BIC clustering
Viterbi resegmentation with energy constraints
Bandwidth and gender identification
c-bic
Agglomerative SID clustering
c-sid
SAD post-filtering
p-asr

Feature Extraction

- MFCC
- The 38 dimensional feature vector consists of 12 cepstrum coefficients, 12 delta and 12 delta-delta coefficients plus the delta and delta-delta energy.
- This set is used in all steps of the **c-std** system, except for the segmentation into small segments where only the static features are used.

Speech Activity Detection

- Speech is extracted from the signal with a Viterbi decoding using Gaussian Mixture Models (GMM) for speech, speech over music, music, silence and noise.

- The aim of the SAD is to remove only long regions without speech such as silence, music, and noise.

Chopping Into Small Segments

- Segmentation of the signal 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 are modeled with a single diagonal Gaussian, i.e., $s_1 \sim N(\mu_1, \Sigma_1)$ and $s_2 \sim N(\mu_2, \Sigma_2)$ with Σ_1 and Σ_2 diagonal.

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

Iterative GMM Segmentation/Clustering Procedure

- Each initial segment is used to seed one cluster, and a GMM with 8 Gaussians and a diagonal covariance matrix is trained by maximum likelihood estimation (MLE) on the segment data.

- Given a sequence of N non-overlapping segments $(s_1, \ldots, s_N)$ with their associated segment cluster labels $(c_1, \ldots, c_N)$, where $c_i \in [1, K]$ and $K \leq N$.

Iterative GMM Segmentation/Clustering Procedure

- The objective function used is a penalized log-likelihood of the form:

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

- where $f(s_i \mid M_{c_i})$ 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.

Viterbi Resegmentation

- The segment boundaries are refined using the last set of GMMs and an additional relative energy-based boundary, within a 1 second interval.

- This is done to locate the segment boundaries at silence portions, so as to avoid cutting words.

Bandwidth and Gender Labeling

- Band (studio or telephone) and gender (male or female) labeling is performed on the segments using 4 GMMs with 64 diagonal covariance matrices, trained on a subset of the 1996/1997 Broadcast News data.

BIC Clustering (1/3)

- 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.
- At each iteration, the two nearest clusters are merged until the stopping criterion is reached.

BIC Clustering (2/3)

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

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

- 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 .

BIC Clustering (3/3)

- The penalty P is

$$P = \frac{1}{2}\left(d + \frac{1}{2}d\left(d+1\right)\right)\log n$$

- where d is the dimension of the feature vector space.
- The merging criterion is that two clusters should be merged if $\Delta BIC < 0$

SID Clustering (1/4)

- After several iterations, 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.

SID Clustering (2/4)

- The feature vector consists of 15 Mel frequency cepstral coefficients plus delta coefficients and delta energy.

- For each gender and channel condition (studio, telephone) combination, **a universal background model** (**UBM**) with 128 diagonal Gaussians is trained on the 1996/1997 English broadcast news data.

- Agglomerative clustering is performed separately for each gender and bandwidth condition, using a **cross log-likelihood ratio**.

SID Clustering (3/4)

- 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 δ is defined as

$$S(c_i, c_j) = \frac{1}{n_i} \log \frac{f\left(x_i \mid M_j\right)}{f\left(x_i \mid B\right)} + \frac{1}{n_j} \log \frac{f\left(x_j \mid M_i\right)}{f\left(x_j \mid B\right)}$$

- where $f(\cdot \mid M)$ is the likelihood of the acoustic frames given the model M, and n_i is the number of frames in cluster c_i.

SID Clustering (4/4)

- The clustering stops when the cross log-likelihood ratio between all clusters is below a given threshold δ optimized on the development data.

SAD Post-Filtering

- In order to filter out short-duration silence segments that were not removed in the initial speech detection step.

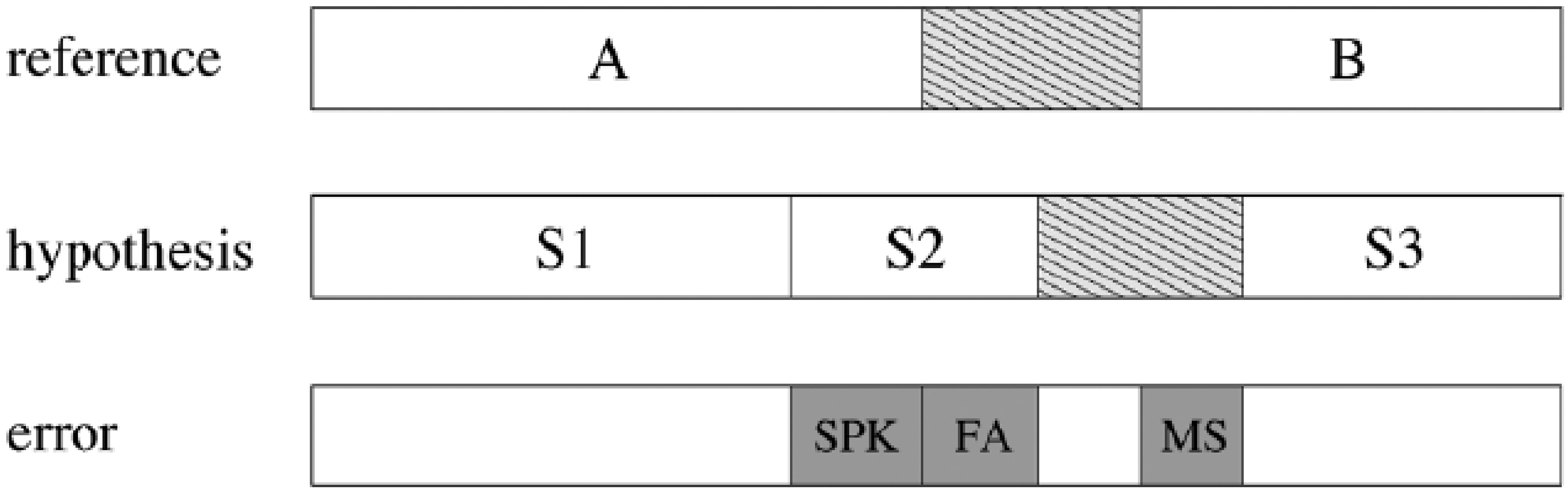
EXPERIMENTS

- ***Development data***: 6 shows of about 30 minutes each, recorded in February 2001 (sources: ABC, CNN, NBC, PRI, VOA), referred to as ’**dev1**’, and 6 shows each of about 30 minutes, recorded in November and December 2003 (sources: ABC, CNBC, CNN, C-SPAN, PBS), referred to as ’**dev2**’;
- ***evaluation (test) data***: 12 shows lasting about 30 minutes, recorded in Dec. 2003 (sources: ABC, CNBC, CNN, CSPAN, PBS, WB17).

EXPERIMENTS

- **Cluster purity** is defined as the ratio between the number of frames by the dominating speaker in a cluster and the total number of frames in the cluster.
- **Cluster coverage** accounts for the dispersion of a given speaker's data across clusters.

DER



DER = Speaker Error (SPK) + False Alarm Speech (FA) + Missed Speech (MS)

EXPERIMENTAL RESULTS

system	*cluster purity (%)*	*coverage (%)*	*overall error (%)*
c-std ($\alpha = \beta = 160$)	95.0	71.6	32.3
c-std ($\alpha = \beta = 230$)	90.6	82.1	24.8
c-bic ($\lambda = 5.5$)	97.1	90.2	13.2
c-sid ($\lambda = 3.5, \delta = 0.1$)	97.9	95.8	7.1

data set	*missed speech (%)*	*false alarm speech (%)*	*speaker error (%)*	*overall error (%)*
dev1	**0.4**	**1.3**	**5.4**	**7.1**
ABC	1.6	1.3	12.4	15.2
VOA	0.3	1.2	2.2	3.7
PRI	0.1	0.9	2.8	3.8
NBC	0.1	1.1	12.0	13.2
CNN	0.5	1.4	5.6	7.6
MNB	0.2	1.8	0.8	2.8
dev2	**0.5**	**3.1**	**4.1**	**7.6**
CSPAN	0.3	2.9	0.1	3.3
CNN	0.6	4.2	5.0	9.8
PBS	0.1	2.8	7.4	10.3
ABC	2.1	6.7	12.5	21.2
CNNHL	0.0	1.4	0.5	1.9
CNBC	0.2	1.0	0.9	2.1

system	*missed speech (%)*	*false alarm speech (%)*	*speaker error (%)*	*overall error (%)*
c-bic	0.4	1.8	14.8	17.0
c-sid	0.4	1.8	6.9	9.1
p-asr	0.6	1.1	6.8	8.5