Speaker Diarization for Conference Room: The UPC RT07s Evaluation System

Jordi Luque, Xavier Anguera and Javier Hernando

TALP Research Center
Universitat Politècnica de Catalunya, UPC

RTs 2007
10-11 May 2007, Baltimore, MD

Outline

Outline

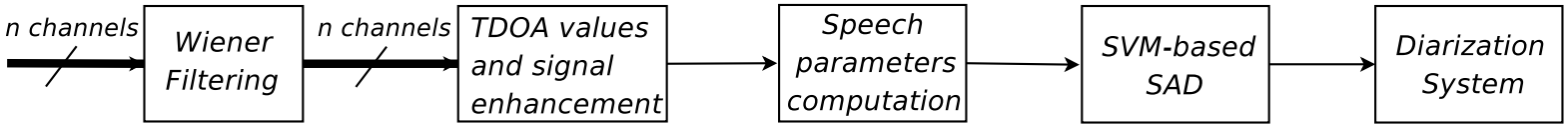
Outline

Outline

Participation objectives

- First participation of the UPC in the Diarization Evaluation
- Consolidation of a baseline system for further research
- Use of the Diarization System from ICSI as baseline
- Changes to the diarization system towards decreasing the runtime while maintaining the performance

Common features with ICSI system

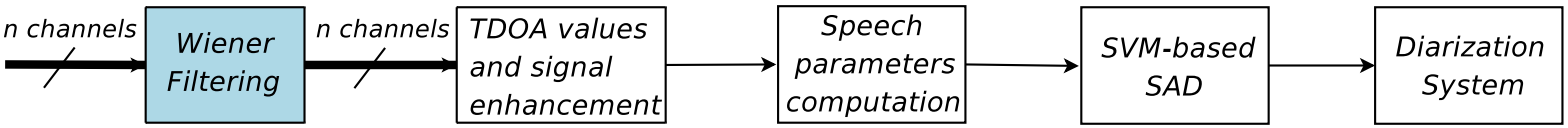


- The sytem is based on a reduced version of the ICSI'06 Diarization system
- Use of the agglomerative system
- Modified BIC criterion to decide when to stop merging clusters
- Linear inicialization of the number of cluster
- Use of the Wiener Filtering and multichannel capabilities from ICSI implementations

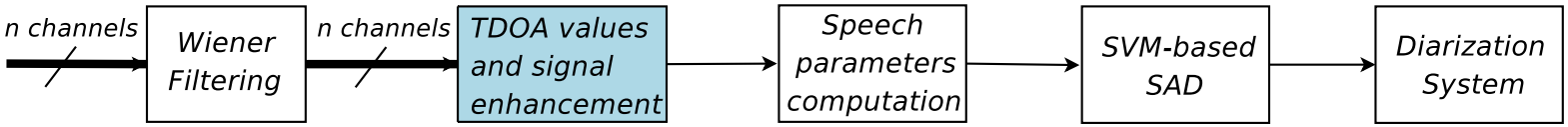
Novelties

- New Speech Activity Detector (SAD) module based on SVM
- New speech parameterization: **Frequency Filtering**
- Changes in the cluster merging in order to avoid small clusters
- Post-processing of the shortest segments at each iteration

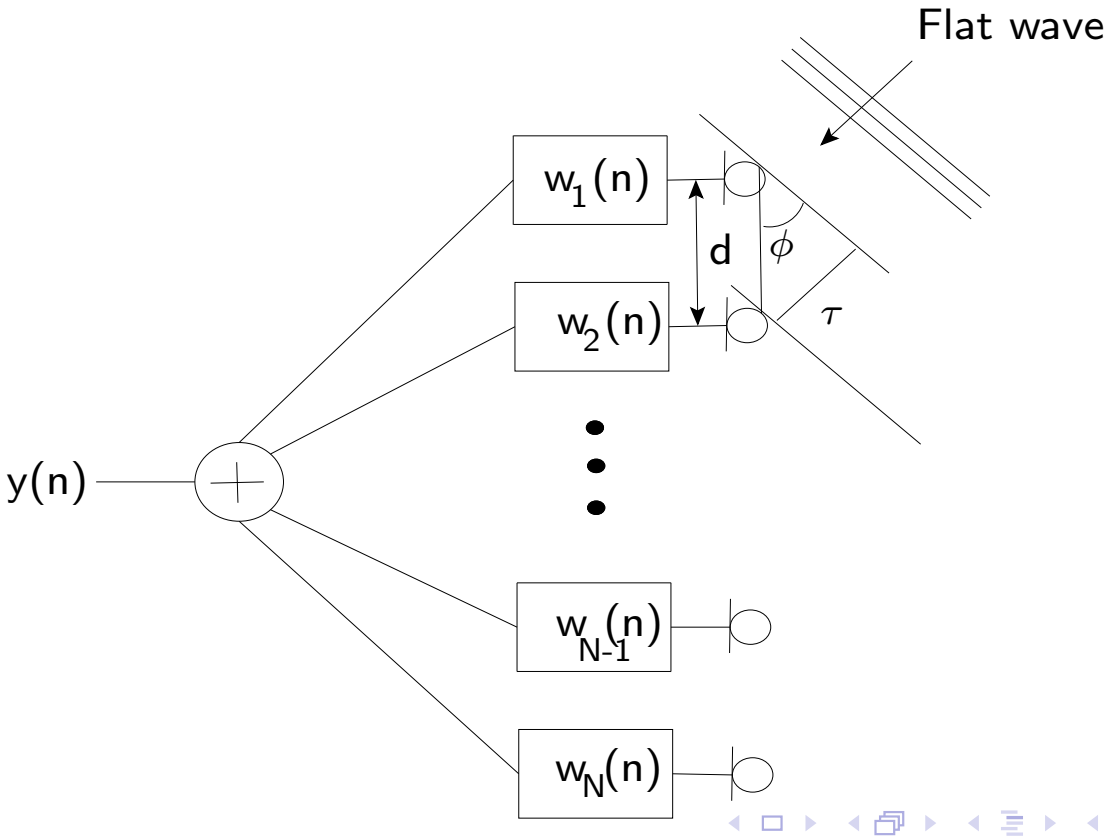
Wiener Filtering



- Use of the ICSI implementation of the Aurora front-end
- Purpose: Avoid stationary noise



Use of the BeamformIt 2.0 from Xavier Anguera



Acoustic Beamforming: Delay and Sum (II)

- We have used a window of 500 ms at a rate of 250 ms
- And all the avalaible channels

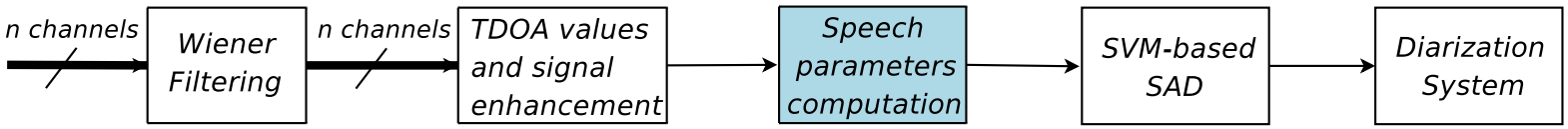
$$y(n) = x_0[n] + \sum_{i=1}^{N-1} W_i x_i[n - d(0,i)]$$

- Estimation of the Time Delay Of Arrival (**TDOA) through the (GCC-PHAT)**

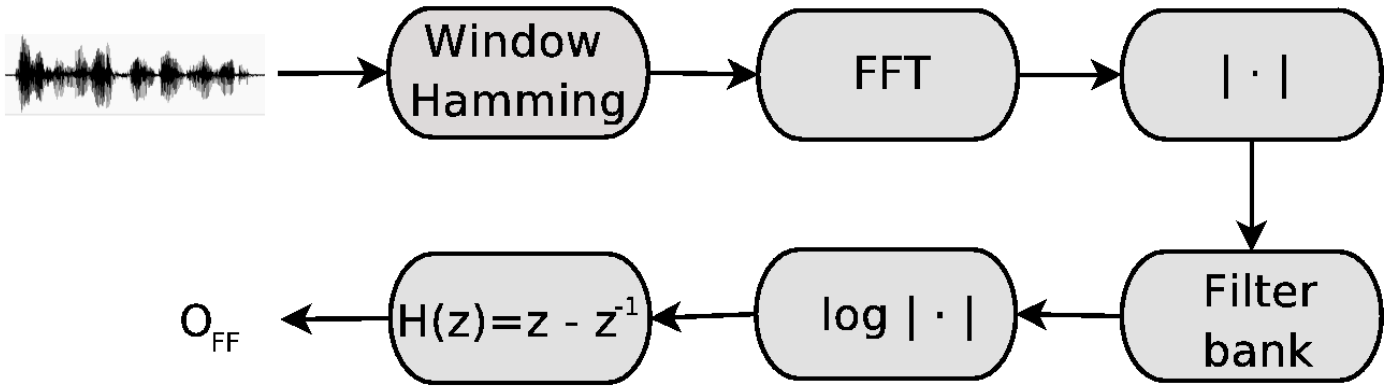
$$G_{PHAT}(f) = \frac{X_i(f)\left[X_j(f)\right]^*}{\left|X_i(f)\left[X_j(f)\right]^*\right|}$$

$$\hat{d}_{ij} = \arg\max_d \hat{R}_{PHAT}(d_{ij})$$

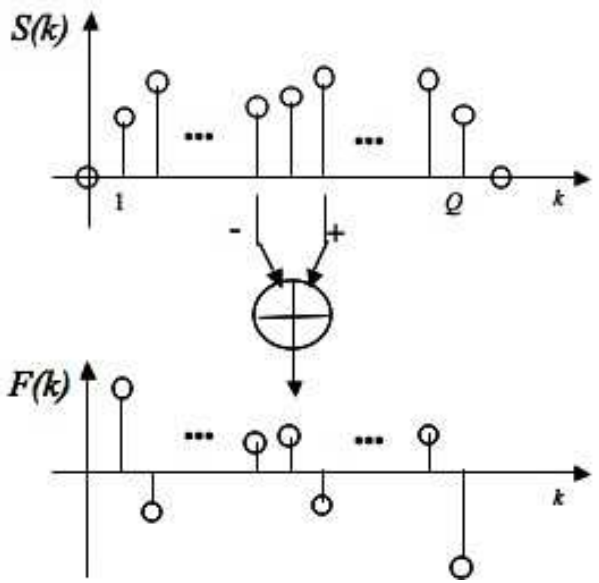
Speech Parameterization: Frequency Filtering



- Computation of Frequency Filtering (FF) parameterization
  - Average of 30 overlapped triangular filters
  - 30 FF coefficients

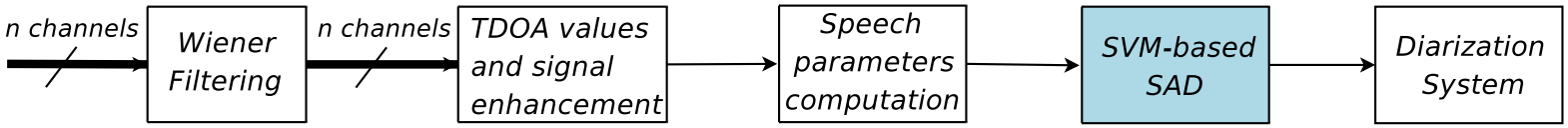


Acoustic System: Baseline System FF parameterization



- Computationally simpler than MFCC
- Compact and uncorrelated
- Frequency meaning, which permits masking, noise subtraction ...
- Have been shown competitive with conventional MFCC

Speech Activity Detection

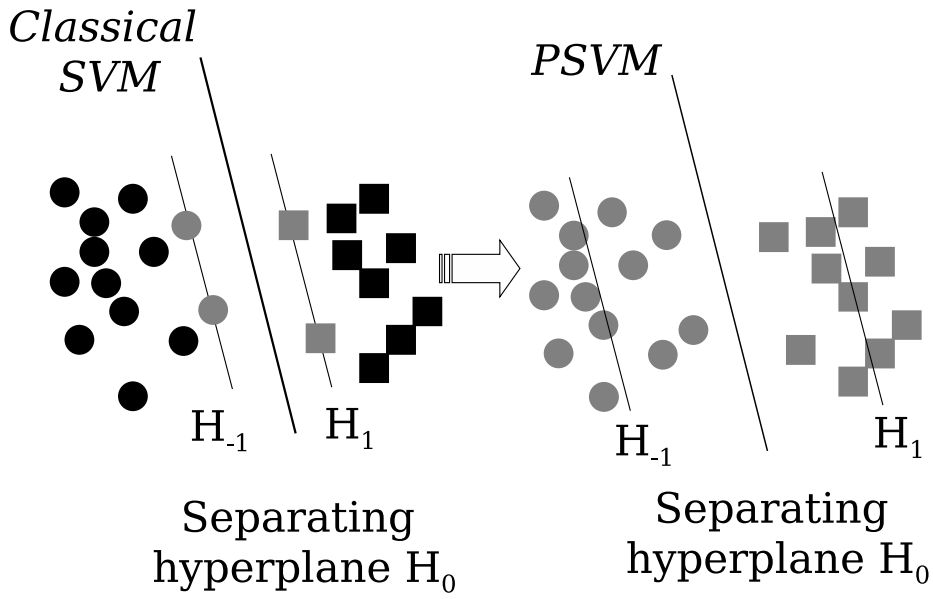


- SAD based on a Support Vector Machine (SVM)
- Two specific modifications in order to adapt to the Evaluation Metrics:
  - NIST = Duration of Incorrect Decissions / Duration of all Speech
  - Missed Spkr = Missed Speech / Duration of All Speech
  - False Alarm = Missed Non-Speech / Duration of All Speech
- Penalize more the Speech class (as NIST metric does) by introducing different costs for the two classes

Speech Activity Detection

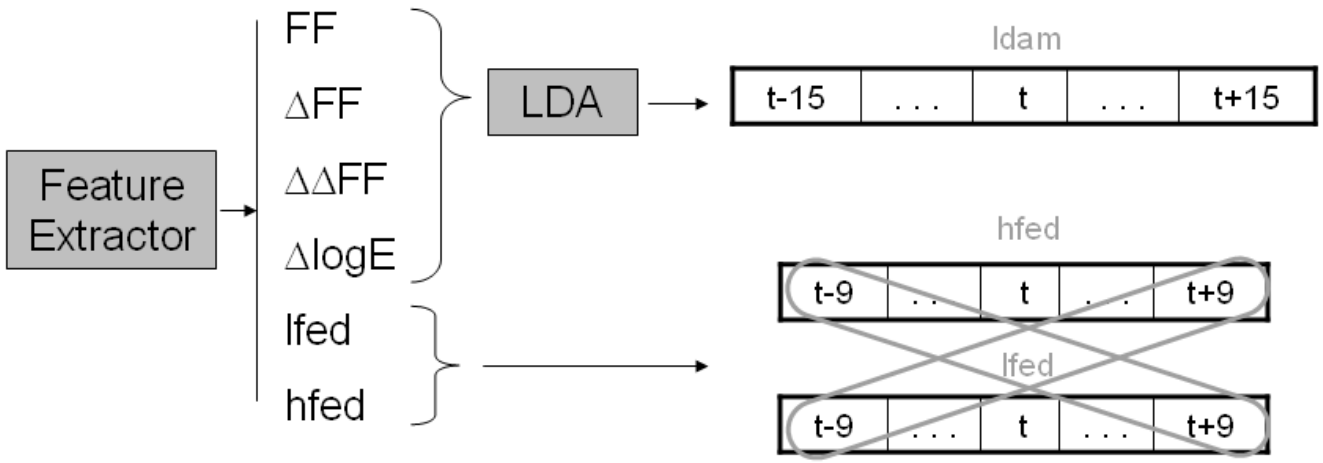
- Dataset reduction (several hundreds of thousands) using an efficient sample selection.
- Main idea: Relaxing the hyperplane condition between two classes

$$y(wx + b) \geq 1 \qquad y(wx + b) = 1$$



Speech Activity Detection: Speech Features

- $16FF + 16\Delta + 16\Delta\Delta + \Delta E = 49$ reduced to a single scalar measure by LDA
- High, low and cross frequency spectral components (focus on the dynamics of the signal along the time)



$$xfed(t) = 1/2 * ([hfed(t-9)*lfed(t+9)]^{1/2} + [hfed(t+9)*lfed(t-9)]^{1/2})$$

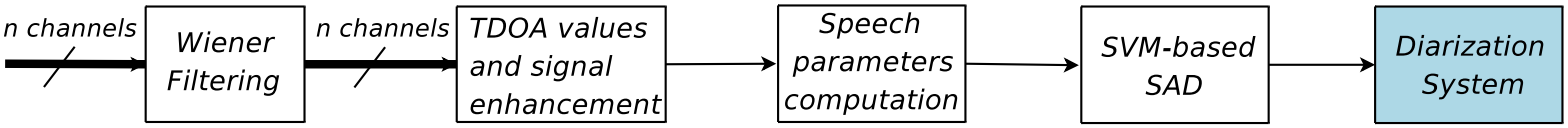
Speech Activity Detection: SAD results

SAD SVM-based (sdm)		
RT'05	RT'06	RT'07
8.03 %	4.88 %	7.03 %

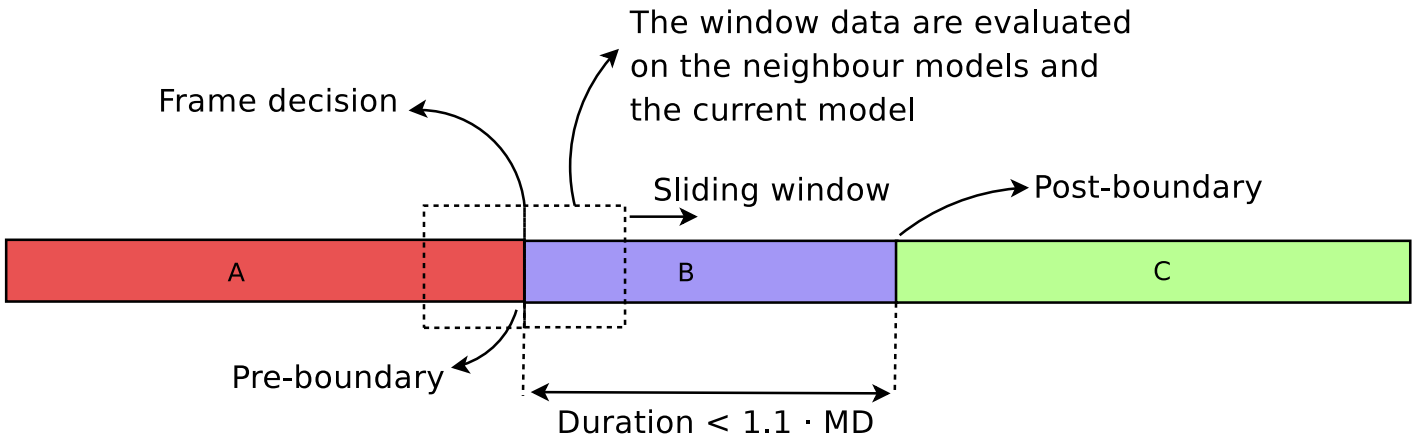
SAD SVM-based (mdm)	
RT'07 mdm-softsad	RT'07 mdm-hardsad
5.39 %	4.72 %

A 12% of relative improvement of the DER (SPKR as SAD) between the two systems submitted

Shortest Segments Post-processing



- Changes in the Complexity Selection algorithm: All clusters modelled with 4 or less Gaussians are rejected
- All those segments with a duration smaller than $1.1 * MD$ are processed by a sliding window
- This kind of segments, usually are associated to false alarm
- The data are splitted between the adjacent clusters



Experimental Set-up and Results

- Evaluation Data from RT'05 used for the training of the SAD classes
- Evaluation Data from RT'06 used for tune the Beamforming, Parameterization and Diarization system parameters

Non-Overlap SPKR Error		
sdm	mdm-softsad	mdm-hardsad
25.06 %	19.65 %	19.75 %

Overlap SPKR Error, **Primary Metric**		
sdm	mdm-softsad	mdm-hardsad
27.72 %	22.70 %	22.59 %

Conclusions

- Novelties:
  - The use of the Beamforming in the MDM condition improves the results obtained in the SDM.
  - The SAD fine ajustment does not imply significant differences in the DER of the whole system
  - Frequency Filtering parameters have obtained better results than the MFCC
  - Post-Processing improves the DER over $1 - 2\%$
- Evaluation:
  - Expensive tuning of the parameters of the system
  - High variance in the DER between different shows i.e, 57.58 DER from CMU show and 5.62 from NIST show

Thanks!

Thank you for your attention!

Questions?