ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

Universidad del País Vasco
Euskal Herriko Unibertsitatea

GTTS Tecnologías Software
Grupo de Trabajo en Tecnologías de Software

An Online Speaker Tracking System for Ambient Intelligence Environments

Maider Zamalloa[1,2], Mikel Peñagarikano[1], Luis Javier Rodriguez-Fuentes[1], Germán Bordel[1], Juan Pedro Uribe[2]

[1]GTTS, Electricity and Electronics Department, University of the Basque Country, Spain
[2]Ikerlan – Technological Research Centre, Spain

ikerlan
ik4 research alliance

Outline

- Introduction
  - The Ambient Intelligence vision
  - Speaker Tracking
- Low-latency Online Speaker Tracking System
- Experiments
- Conclusions

Introduction – The AmI vision

- *Ubiquitous Computing*
  - Envisages the integration of computing and telecommunication capabilities in daily objects
  - A term defined by M. Weiser in 1991:

    … The most profound technologies are those that dissapear. They wave themselves into the fabric of everyday life until they are indistinguishable from it….

- The ***Ambient Intelligence*** **(AmI) vision**
  - Generalizes the Ubiquitous Computing term
  - A vision oriented towards the usability of ubiquitous technologies and promoted by the group ISTAG of the European Commission
  - It was defined in 2001 through a set of scenarios and recommendations

Introduction – The AmI vision

- The AmI paradigm is caracterized by systems that are:
  - *Embedded*: Integrated into the environment
  - *Context-aware*: Recognize users and user situational context
  - *Personalized*: Tailored to user needs
  - *Adaptive*: Change in response to user
  - *Anticipatory*: Anticipate to user needs
- Main objective: support people carrying out everyday life activities in a *natural* way
- Transparency is critical

Introduction – The AmI vision

- The AmI paradigm is caracterized by systems that are:
  - *Embedded*: Integrated into the environment
  - *Context-aware*: Recognize users and user situational context
  - *Personalized*: Tailored to user needs
  - *Adaptive*: Change in response to user
  - *Anticipatory*: Anticipate to user needs
- Main objective: support people carrying out everyday life activities in a *natural* way
- Transparency is critical →

Natural and Intelligent Interfaces are needed

Introduction – Speaker Tracking

- Speech is a natural interface for human interaction
  - It conveys many user related information:
    - The message
    - The language of the message
    - The speaker location
    - The speaker identity
    - The emotional state of speaker
    - etc.
- It is a very suitable means to support user interaction, adaptation and monitorization
- Speaker tracking and speaker diarization technologies may be used

Introduction – Speaker Tracking

- In Speech Technologies area, speaker diarization and speaker tracking are well known tasks
- Both answer the question: Who spokes when?
- But differ in:
  - Speaker Tracking aims to detect audio segments correspondiing to a known set of target speakers
  - Speaker Diarization consists of detecting speaker turns without any prior knowledge about the target speakers

Introduction – Speaker Tracking

GTTS Tecnologías Software

- Speaker tracking and diarization primary application domains
  - Telephone conversations
  - Broadcast news
  - Meeting recordings
- Common approaches consists of two uncoupled steps:
  - Audio Segmentation
  - Speaker detection
- In an AmI Environment speaker detection must be continuous and real-time

Introduction – Speaker Tracking

- Speaker tracking and diarization primary application domains
  - Telephone conversations
  - Broadcast news
  - Meeting recordings

Audio recording is fully available before processing!!

- Common approaches consists of two uncoupled steps:
  - Audio Segmentation
  - Speaker detection
- In an AmI Environment speaker detection must be continuous and real-time

Introduction – Speaker Tracking

- Speaker tracking and diarization primary application domains
  - Telephone conversations
  - Broadcast news
  - Meeting recordings

Audio recording is fully available before processing!!

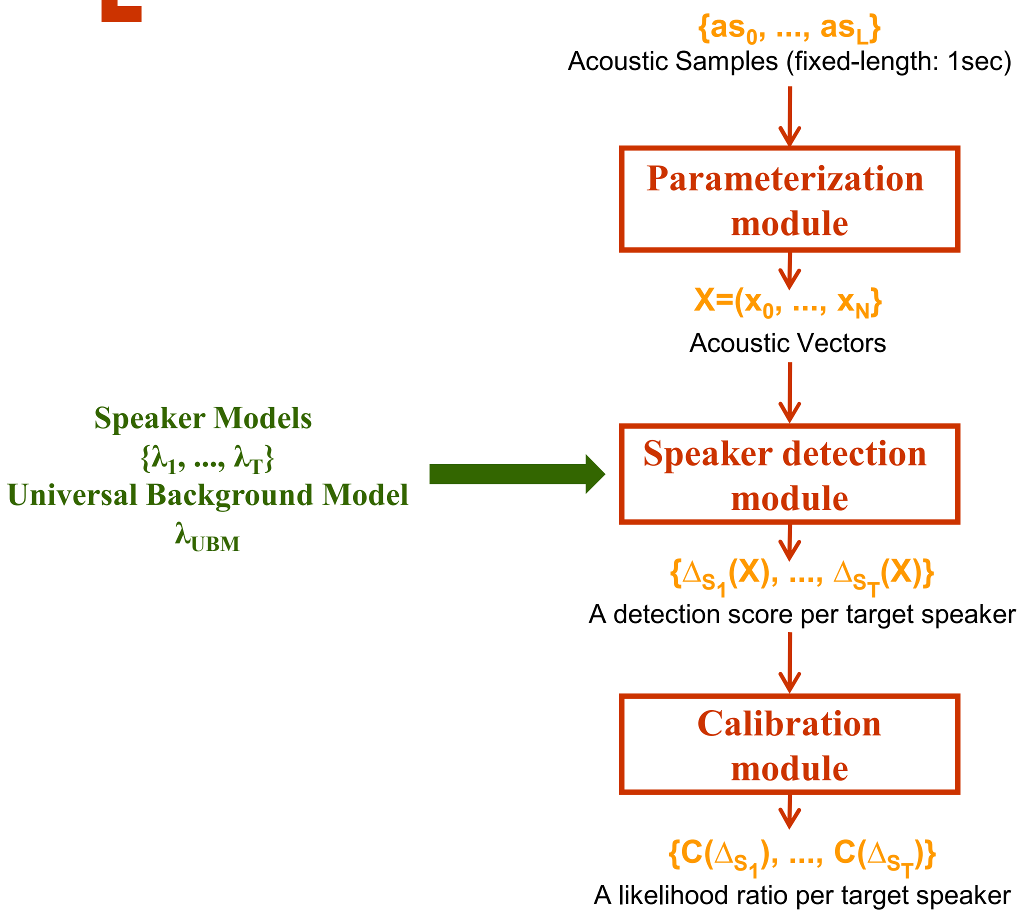
- Common approaches consists of two uncoupled steps:
  - Audio Segmentation
  - Speaker detection
- In an AmI Environment speaker detection must be continuous and real-time

State of the arte approaches are not suitable for low-latency online speaker detection

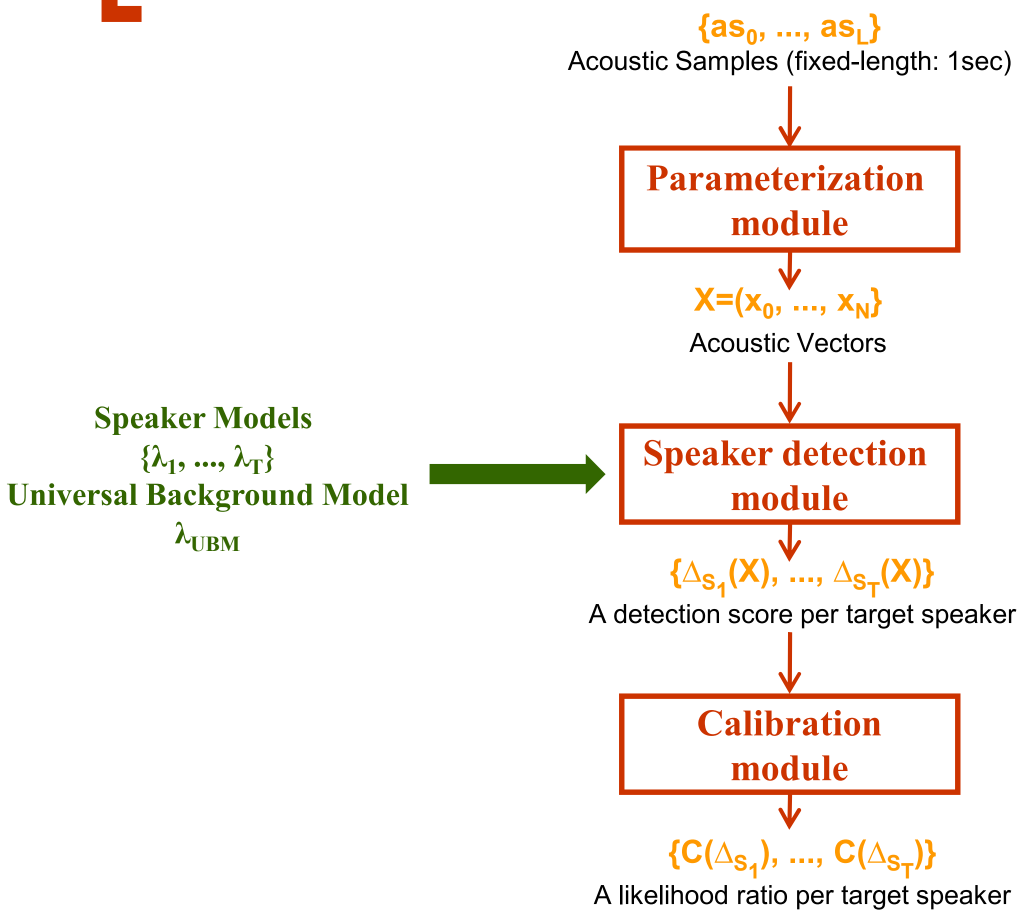
Low-latency Online Speaker Tracking System

- System is designed for an intelligent home environment
  - It tracks known speakers continuously
  - The expected number of targets is low (i.e. the members of a family)
  - The scenario requires almost instantaneous (low-latency) speaker tracking decisions
- So, a very simple speaker tracking algorithm is designed
  - Joint speaker segmentation and speaker detection is performed
  - Fixed-length audio segments are defined and processed

Low-latency Online Speaker Tracking System



Low-latency Online Speaker Tracking System



$$\text{Decision} = \begin{cases} S_t & \text{if } \max_{1 \le t \le T}\{C(\Delta_{S_t})\} \ge \theta \\ \text{impostor} & \text{otherwise} \end{cases}$$

Low-latency Online Speaker Tracking System

- Parameterization Module
  - Channel Normalization: Dynamic Cepstral Mean Normalization
  - Acoustic Vectors: 12 Mel Frequency Cepstral Coefficients (MFCC) and deltas
  - Parameterization is done by Sautrela Framework (***Penagarikano, www.sautrela.org***)
- Speaker Detection Module
  - Acoustic Speaker Models $\lambda_t \in [\lambda_1, \dots, \lambda_T]$
    - A Gaussian Mixture Model (GMM) adapted from an universal model λ_{UBM}
    - In adaptation, non-overlapped single-speakers segments are used
  - Given λ_t and the parameterized acoustic segment ***X***, the speaker detection score $\Delta_{S_t}(X)$ is:
    - $\Delta_{S_t}(X) = L(X|\lambda_t) - L(X|\lambda_{UBM})$ where $L(X|\lambda)$ is the log-likelihood of ***X*** given λ

M. Penagarikano and G. Bordel, "SAUTRELA: A Highly Modular Open Source Speech Recognition Framework", In Proceedings of the IEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2005.

Low-latency Online Speaker Tracking System

- Calibration Module
  - Maps detection scores to likelihood ratios by applying a linear transform *C*:
    - $\{\Delta_{S_t} \mid t \in [1,T]\} \rightarrow \{C(\Delta_{S_t}) \mid t \in [1,T]\}$
  - Scaling parameters are computed over a development corpus
    - Optimization process is based on *Maximizing Mutual Information*
  - Minimum expected cost based decision threshold is applied over calibrated scores
    - $Thresold = \theta = ln\left(\frac{C_{fa}(1-P_{target})}{C_{miss}P_{target}}\right)$
    - $Decision = \begin{cases} S_t & if \ \max_{1 \le t \le T}\{C(\Delta_{S_t})\} \ge \theta \\ impostor & otherwise \end{cases}$
  - Calibration is done by FoCal toolkit (***Brummer, sites.google.com/site/nikobrummer/focal***)

Experimental setup

- AMI (*Augmented Multipart Interaction*) Corpus
  - Real-time human interaction in the context of smart meeting rooms
  - Audio & video data collected in 3 instrumented rooms (Edinburgh, IDIAP, TNO)
  - 4 english (mostly non-native) speakers per meeting; 4 meetings per session; 30 minutes meetings
- Experiments are based on 15 Edinburgh sessions
  - 3 speakers act as target, the fourth one as impostor
  - Two independent subsets are defined:
    - Development (Dev) : 8 sessions (32 meetings)
    - Evaluation (Eval) : 7 sessions (28 meetings)
  - Dev and Eval sets consist of:
    - Train dataset: 2 meetings per session (random selection)
    - Test dataset: 2 meetings per session
  - For time references AMI corpus manual annotations are used

Experimental setup

- Two online speaker tracking systems which differ in UBM estimation data:
  - **UBM-g** uses15 gender-balanced meetings from all sites except Edinburgh
  - **UMB-t** uses only speech data from target speakers
- System performance is compared to an offline reference system following a clasical two-stage approach
  - Audio segmentation is done by a similar approach to well known BIC
  - Speaker detection is carried out by computing speaker model likelihood ratios
- Performance measure:
  - $F_{measure} = \frac{2 \times PRC \times RCL}{PRC + RCL}$ ranges from 0 to 1, where:
    - Precision (PRC) computes correctly detected target time from total target time
    - Recall (RCL) estimates correctly detected target time from actual target time

Results – online vs offline

- The expected performance loss of the low-latency online system is low:

		Dev		
		PRC	RCL	$F_{measure}$
UBM-g	online	0.66	0.92	0.77
	ref	0.67	0.93	0.78
UBM-t	online	0.67	0.91	0.77
	ref	0.69	0.92	0.79

Results – online vs offline

- The expected performance loss of the low-latency online system is low:

		Dev			Eval		
		PRC	RCL	$F_{measure}$	PRC	RCL	$F_{measure}$
UBM-g	online	0.66	0.92	0.77	0.69	0.92	0.78
	ref	0.67	0.93	0.78	0.69	0.93	0.79
UBM-t	online	0.67	0.91	0.77	0.71	0.91	0.8
	ref	0.69	0.92	0.79	0.72	0.92	0.81

Results – online vs offline

- The expected performance loss of the low-latency online system is low:

		Dev			Eval		
		PRC	RCL	$F_{measure}$	PRC	RCL	$F_{measure}$
UBM-g	online	0.66	0.92	0.77	0.69	0.92	0.78
	ref	0.67	0.93	0.78	0.69	0.93	0.79
UBM-t	online	0.67	0.91	0.77	0.71	0.91	0.8
	ref	0.69	0.92	0.79	0.72	0.92	0.81

With respecto to the classical offline system:
UBM-g: 1.26% relative degradation
UBM-t: 1.23% relative degradation

Results – UBM-g vs UBM-t

- UBM-t system slightly outperforms the performance of UBM-g system:

		Dev			Eval		
		PRC	RCL	$F_{measure}$	PRC	RCL	$F_{measure}$
online	UBM-g	0.66	0.92	0.77	0.69	0.92	0.78
	UBM-t	0.67	0.91	0.77	0.71	0.91	0.8
reference	UBM-g	0.67	0.93	0.78	0.69	0.93	0.79
	UBM-t	0.69	0.92	0.79	0.72	0.92	0.81

Results – UBM-t vs UBM-g

- UBM-t system slightly outperforms the performance of UBM-g system:

		Dev			Eval		
		PRC	RCL	$F_{measure}$	PRC	RCL	$F_{measure}$
online	UBM-g	0.66	0.92	0.77	0.69	0.92	0.78
	UBM-t	0.67	0.91	0.77	0.71	0.91	0.8
reference	UBM-g	0.67	0.93	0.78	0.69	0.93	0.79
	UBM-t	0.69	0.92	0.79	0.72	0.92	0.81

Results support the use of a specific UBM for room and speaker set:

👍 There is a high consistency between the UBM and target speakers

👎 But a different UBM model must be estimated for each set of target speakers

Results – Calibration

- Calibration stage leads to a better performance in all cases:

		Uncalibrated			Calibrated		
		PRC	RCL	$F_{measure}$	PRC	RCL	$F_{measure}$
Dev	UBM-g	0.66	0.92	0.77	0.81	0.8	0.81
	UBM-t	0.67	0.91	0.77	0.82	0.83	0.82
Eval	UBM-g	0.69	0.92	0.78	0.78	0.85	0.8
	UBM-t	0.71	0.91	0.8	0.81	0.85	0.83

(Have a look at the paper for the results of the reference system)

Results – Calibration

- Calibration stage leads to a better performance in all cases:

		Uncalibrated			Calibrated		
		PRC	RCL	$F_{measure}$	PRC	RCL	$F_{measure}$
Dev	UBM-g	0.66	0.92	0.77	0.81	0.8	0.81
	UBM-t	0.67	0.91	0.77	0.82	0.83	0.82
Eval	UBM-g	0.69	0.92	0.78	0.78	0.85	0.8
	UBM-t	0.71	0.91	0.8	0.81	0.85	0.83

2.56% relative improvement

(Have a look at the paper for the results of the reference system)

Results – Calibration

- Calibration stage leads to a better performance in all cases:

		Uncalibrated			Calibrated		
		PRC	RCL	$F_{measure}$	PRC	RCL	$F_{measure}$
Dev	UBM-g	0.66	0.92	0.77	0.81	0.8	0.81
	UBM-t	0.67	0.91	0.77	0.82	0.83	0.82
Eval	UBM-g	0.69	0.92	0.78	0.78	0.85	0.8
	UBM-t	0.71	0.91	0.8	0.81	0.85	0.83

(Have a look at the paper for the results of the reference system)

3.75% relative improvement

Conclusions

- A online speaker tracking for an AmI scenario is proposed
  - Processes continuous audio streams
  - Outpus an identification decision for fixed-length segments
- The system performance is compared to a reference system based on offline segmentation
  - Even if speaker tracking actually takes advantage from an offline segmentation, online system presents little degradation
  - Depending on the scenario and required latency, offline segmentation may not be feasible
- Better results are attained when the UBM matches test conditions (same room, same speakers)

Thank you!

Any questions?

ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

Universidad del País Vasco — Euskal Herriko Unibertsitatea

GTTS Tecnologías Software
Grupo de Trabajo en Tecnologías de Software

An Online Speaker Tracking System for Ambient Intelligence Environments

Maider Zamalloa[1,2], Mikel Peñagarikano[1], Luis Javier Rodriguez-Fuentes[1], Germán Bordel[1], Juan Pedro Uribe[2]

[1]GTTS, Electricity and Electronics Department, University of the Basque Country, Spain
[2]Ikerlan – Technological Research Centre, Spain

ikerlan
ik4 research alliance