

Multiple Speaker Detection using Energy, Spectral and Voicing-related Features

**Jürgen Geiger, Ravichander Vipperla, Simon Bozonnet,
Nicholas Evans, Gerhard Rigoll**
(**EURECOM, TUM**)

Category: Segmentation (S)

Task: SES-2 multiple speaker detection

Acknowledgements:



Outline

- **Problem of overlapping speech**
- **System description**
- **Experiments**
- **Summary and conclusions**

PROBLEM OF OVERLAPPING SPEECH

Problem of overlapping speech

- **Overlapping speech:**
 - Two or more speakers speaking at the same time
 - Back-channel, approbation/opposition, early start, voluntary jamming
 - **Influence on Speaker Diarization: 22 % of error (DER) due to overlap [1]**
 - Impure speaker models
 - Missed speakers during overlap segments
- ➡ Big potential for improvement of DER
- **Also interesting for other tasks, e.g. ASR**

[1] Huijbregts et al., "The Blame Game: Performance Analysis of Speaker Diarization System Components", Interspeech'07

Related Work

- **3-class HMM system using multiple features [2]**
 - Classes speech, nonspeech, overlap
 - MFCCs + other spectral features
- **Other recent work uses similar systems**
 - Prosodic features, HMM [3]
 - Spatial features, HMM [4]
- **Other approach:**
 - Convolutional Non-negative Sparse Coding, frame-level classifier [5]
 - CNSC with embedded HMM framework [6]

[2] Boakye et al., "Overlapped Speech Detection for Improved Diarization in Multi-Party Meetings," ICASSP'08

[3] Zelenak et al., "Overlap Detection for Speaker Diarization by Fusing Spectral and Spatial Features", Interspeech'10

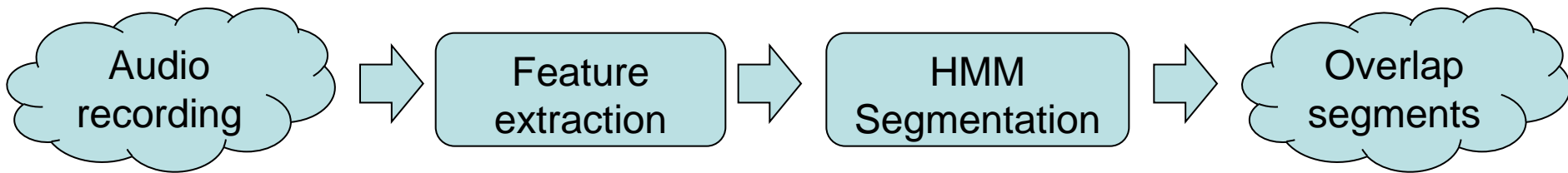
[4] Zelenak et al., "The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization", Interspeech'11*

[5] Vipplerla et al., "Speech Overlap Detection and Attribution Using Convolutional Non-Negative Sparse Coding", ICASSP'12

[6] Geiger et al., "Convolutional Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization", Interspeech'12

SYSTEM DESCRIPTION

System overview



Feature extraction

■ Energy, spectral and voicing-related features

- 22 features + delta coefficients
- Frame rate: 20 ms
- Window size: 25 or 60ms
(depending on the features)
- No CNSC features! (different from [6])

■ Feature selection

- using AMI meeting corpus [7]
- Based on KL divergence
(ovlp/all frames)

■ Feature standardization

List of features

MFCC 1-12

Loudness (auditory model based)

Energy in band 250 – 50 Hz

Energy in band 1 kHz – 4 kHz

Spectral flux

Spectral kurtosis

Spectral harmonicity

F0 (subharmonic summation)

Probability of voicing

Jitter

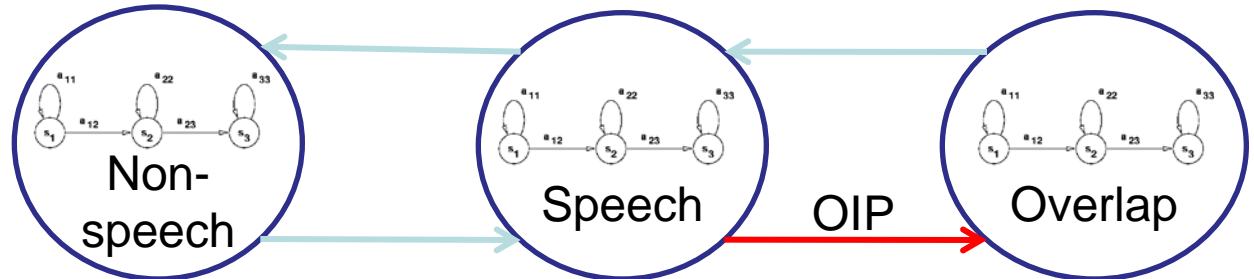
Shimmer(local)

[6] Geiger et al., “Convolutional Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization”, Interspeech’12

HMM Segmentation

- Hidden Markov Model system
- 3 classes (models), each with 3 states

- Nonspeech
- speech
- Overlap



- Training: Iterative mixture-splitting with successive EM re-estimation
- Viterbi Decoding
- Overlap Insertion Penalty OIP
 - **Penalty** for detecting overlap in the HMM word net
 - Trade-off between precision and recall

EXPERIMENTS

Experimental Settings

- **Database: ETAPE 2011 data set**
 - 30 hours of broadcast shows
 - Divided in training, development and test
 - Dev: 5% nonspeech, 87% speech, 8% overlap
- **For test set experiments: trained with train+devel sets**
- **HMM system parameters:**
 - # mixtures:
 - ☞ nonspeech and overlap 64,
 - ☞ speech: different setups from: 64 to 256 components
 - OIP: 0, -10, -50, -100

Evaluation Metrics

Averaged frame-level metrics:

- **Precision (P)**

- correctly detected overlap frames / detected overlap frames

- **Recall (R)**

- correctly detected overlap frames / all overlap frames

- **Interests:**

- High precision: eliminate main of ovlp to train pure models
- High recall: speaker attribution for overlap segments

Results – Development Set

- # mixtures for speech model highly influential
- OIP increased to get higher precision

OIP=0 {		Speech mix	Precision	Recall
		64	17.6	45.8
		128	29.1	26.1
		256	38.1	17.4
Speech mix=256 {		OIP	Precision	Recall
		0	38.1	17.4
		-10	42.1	15.5
		-50	44.8	8.4
		-100	40.2	4.8
		AMI Corpus	61.0	24.0

Precision,
Recall in %
overlap
frames

[6] Geiger et al., “Convolutional Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization”, Interspeech’12

Results – Eval Set

- Training with train + devel data
- 4 setups: 64/128/256 mix, OIP=0 and 128 mix, OIP=-10

Speech mix	OIP	Precision	Recall
64	0	6.8	76.0
128	0	7.9	44.0
128	-10	7.8	39.7
256	0	6.7	21.7

- Too many false alarms

SUMMARY AND CONCLUSIONS

Summary and Conclusions

- **Features: Energy, spectral and voicing features**
 - no CNSC
- **Detection: 3 classes HMM segmentation**
 - trained on dev and train ETAPE datasets
 - Use of Overlap Insertion Penalty to tune the transition speech/ovlp
- **Poor optimization: high differences of performance between dev/eval sets**
 - train different models for different types of shows (TV, radio, etc.)?

Thank you very much!