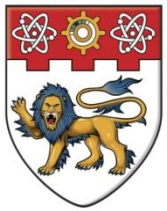


NANYANG TECHNOLOGICAL UNIVERSITY

Overlap Detection in Speaker Diarization: Key to Social Behavior Understanding

presented by

Debsubhra CHAKRABORTY
PhD student
Institute for Media Innovation/ Interdisciplinary Graduate School

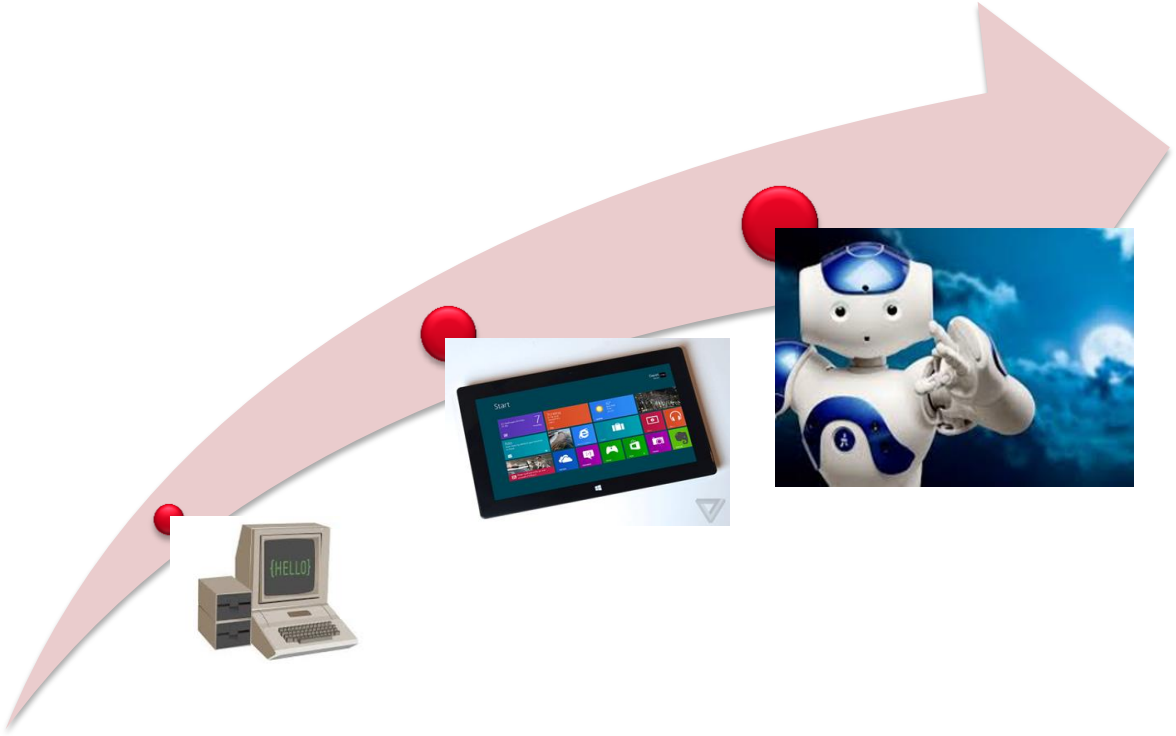
Supervisor: **Assoc. Prof. Justin Dauwels (EEE)**
Co-Supervisor: **Prof. Daniel Thalmann (IMI/SCE)**

25th October, 2016

Outline

- Introduction
- Literature Survey
- Audiobook Corpus
- Results
- Conclusion

Introduction: Social Robotics

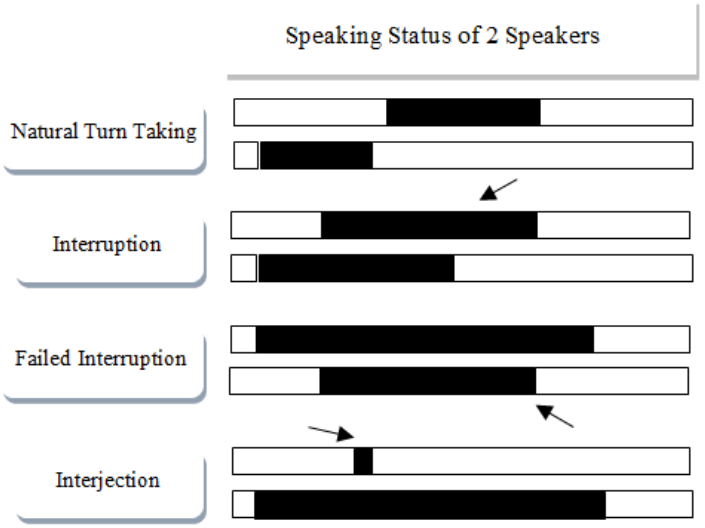


- Computers from tools to facilitating human-human interaction to social robotics
- Interest in automatic analysis of human behavior

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). Human computing and machine understanding of human behavior: a survey. In *Artifical Intelligence for Human Computing* (pp. 47-71). Springer Berlin Heidelberg.

Introduction: Non-verbal Cues

- Conversations contain non-verbal cues
- Audio non-verbal cues – tone, natural turns, interrupts, interjections etc.

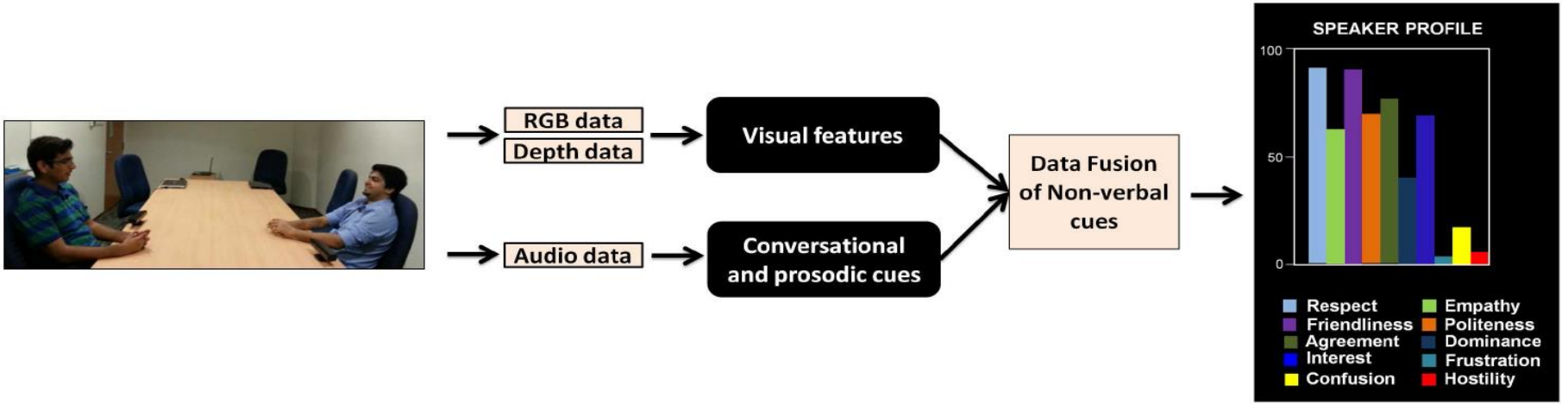


- Social Signal Processing

1. Hassin, R. R., Uleman, J. S., & Bargh, J. A. (Eds.). (2004). *The new unconscious*. Oxford University Press.
2. Knapp, M., Hall, J., & Horgan, T. (2013). *Nonverbal communication in human interaction*. Cengage Learning.
3. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schröder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing.*Affective Computing, IEEE Transactions on*, *3*(1), 69-87. (figure source)

Introduction: Existing System

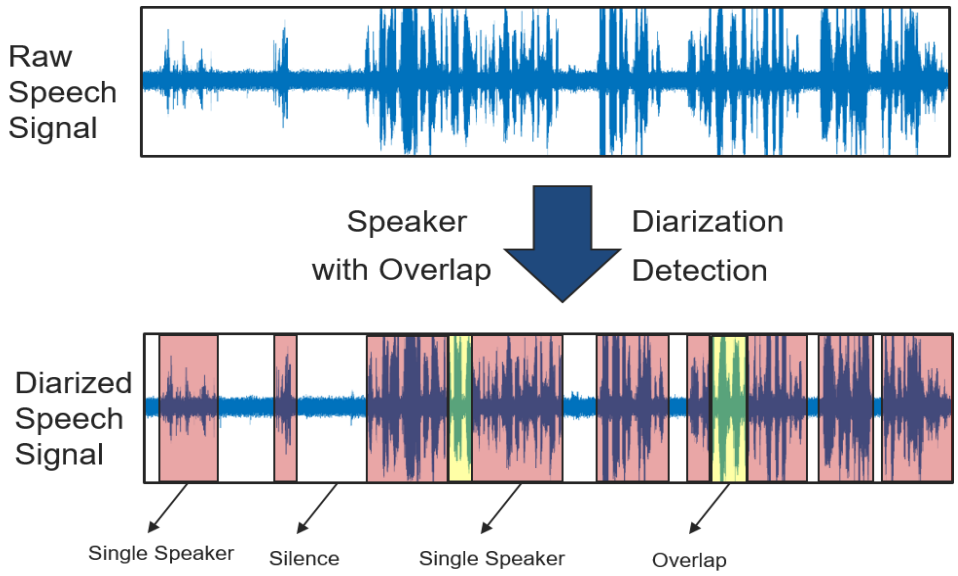
- Applications: Skype interview, online course, defence



- Audio captured through separate microphones
- Easy to detect audio features crucial to social signals

Introduction: Speaker Diarization

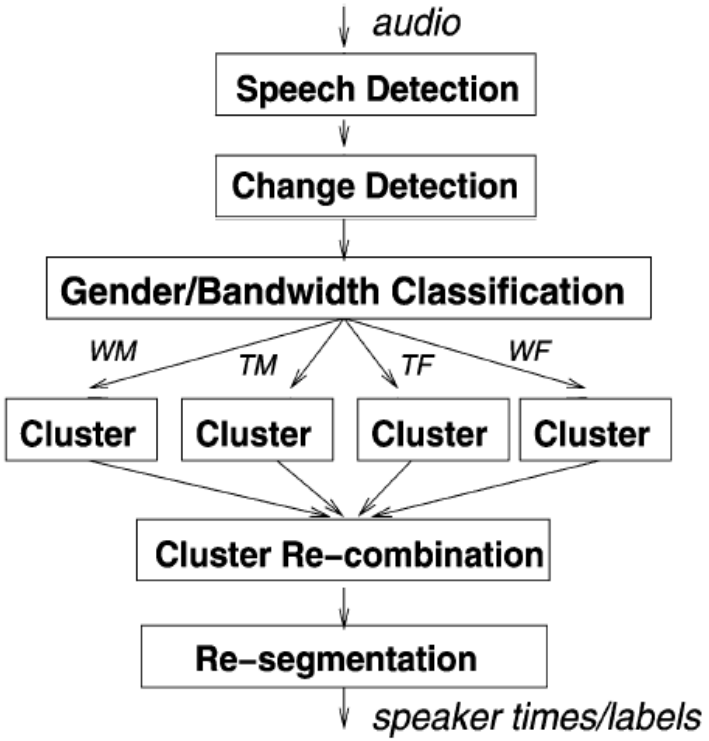
- Diarization is essential aspect of automatic speech recognition systems for single-channel speech
- Diarization: Who spoke when?
- Overlap detection in diarization critical, yet under-addressed
- Overlap detection key to understanding social behavior



Literature Review: Speaker Diarization

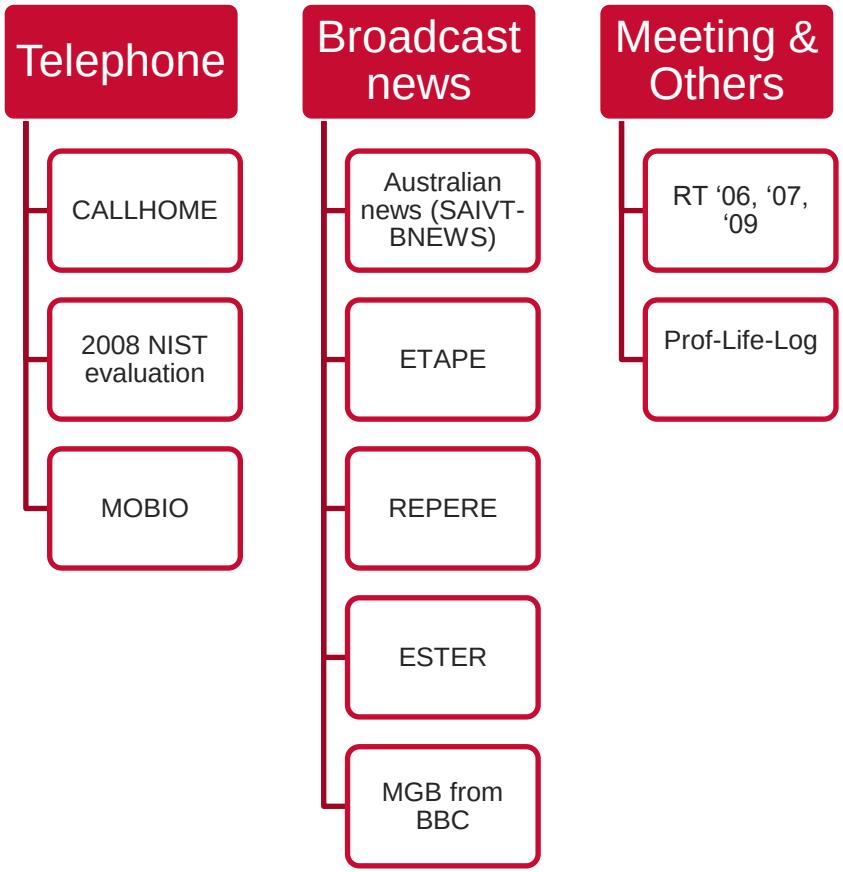
Speaker Diarization

- Distinguish speech – GMM, Viterbi segmentations
- Change detections – calculate distance and compare
- Gender classification – reduces load on clustering
- Clustering – agglomerative
- Recombination – better models
- Resegmentation – refine segment boundaries



Anguera, Xavier, et al. "Speaker diarization: A review of recent research."*IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012): 356-370.

Literature Review: Corpora Used



Literature Review: Methods Used

Feature/ Model

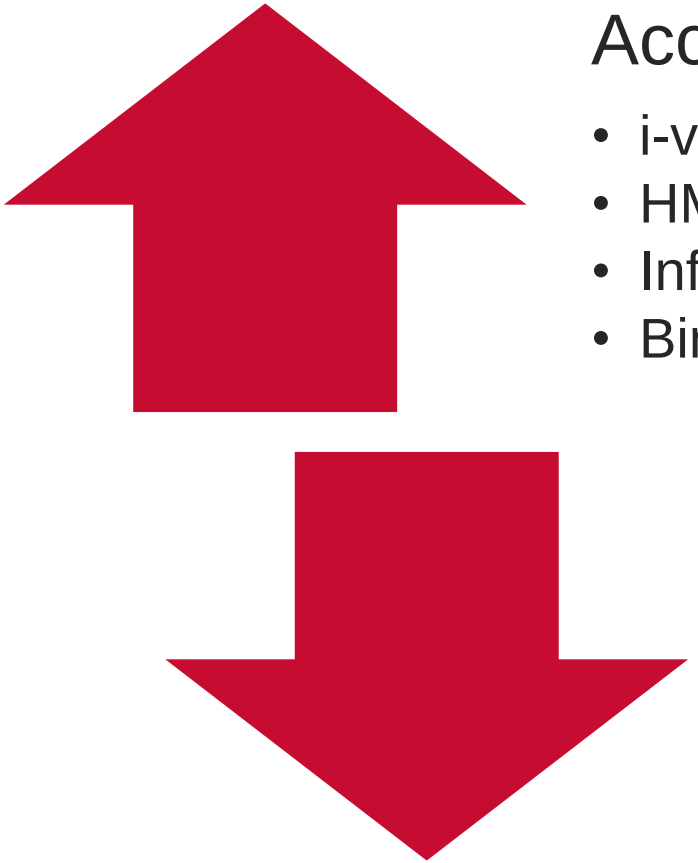
- Enhanced spectrogram to detect overlapped sections (pyknogram)
- CRF framework for AV association of voices and face clusters
- Non-speech as side info for Information Bottleneck Method
- Variational Bayes applied to i-vectors
- KL-HMM as a full diarization technique in IB-based methods
- Binary keys

Resegmentation

- HMM based algorithm on i-vector subspace instead of Viterbi method
- Complete-linkage clustering & pairwise voting

1. Shokouhi, Navid, et al. "Robust overlapped speech detection and its application in word-count estimation for Prof-Life-Log data." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.
2. Paul, Gay, et al. "A conditional random field approach for audio-visual people diarization." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.
3. Yella, Sree Harsha, and Hervé Bourlard. "Information bottleneck based speaker diarization of meetings using non-speech as side information." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.
4. Zheng, Rong, et al. "Variational bayes based i-vector for speaker diarization of telephone conversations." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.
5. Madikeri, Srikanth and Hervé Bourlard. "KL-HMM based speaker diarization system for meetings." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.
6. Delgado, Héctor, Corinne Fredouille, and Javier Serrano. "Towards a complete binary key system for the speaker diarization task."*INTERSPEECH*. 2014.

Literature Review: Speed vs. accuracy



Accuracy

- i-vectors
- HMM-GMM
- Information Bottleneck
- Binary Keys

Speed

- i-vectors
- HMM-GMM
- Information Bottleneck
- Binary Keys

Literature Survey: Motivation

- Current methods mostly use **unsupervised** learning employing Hidden Markov Models modeling Gaussian mixtures
- Overlap areas are difficult to identify due to their short duration, and characteristics somewhere in between 2 speakers
- Most of the methods do not consider overlap at all, and hence there is no dedicated corpus for it
- Our idea:
  - Start with something simple: two speakers with minimum background noise, with overlap sections exactly known
  - Create a large corpus of synthetic conversations where overlap time-periods are exactly known
  - Extract features specific to overlaps and apply various algorithms to reliably detect silence, single-speaker and overlap zones

Audiobook Corpus: LibriSpeech

- This corpus contains data for over 1000 hours
- It is divided into dev-clean, test-clean, train-clean etc. sets
- For now, only dev-clean set has been selected
- Dev-clean contains 40 speakers, 20 Male, 20 Female

Audiobook Corpus: Conversations from audiobooks

- Created 4570 audio files containing conversations of about 3 minutes each
- Total 228 hours of data
- Overlaps can be of 1-1.5 seconds and pauses of 0.5-1 seconds
- Each overlap or pause is placed at the end of a speaker turn
- Roughly in 65% of cases it is an overlap
- Timings are simultaneously noted in excel file

Results: Features

- Feature set is a 26-column matrix

MFCC

- 12 coefficients + their 1st derivatives
- 60 ms window, advanced by 20 ms
- Centralized by subtracting mean from each column

RMS energy

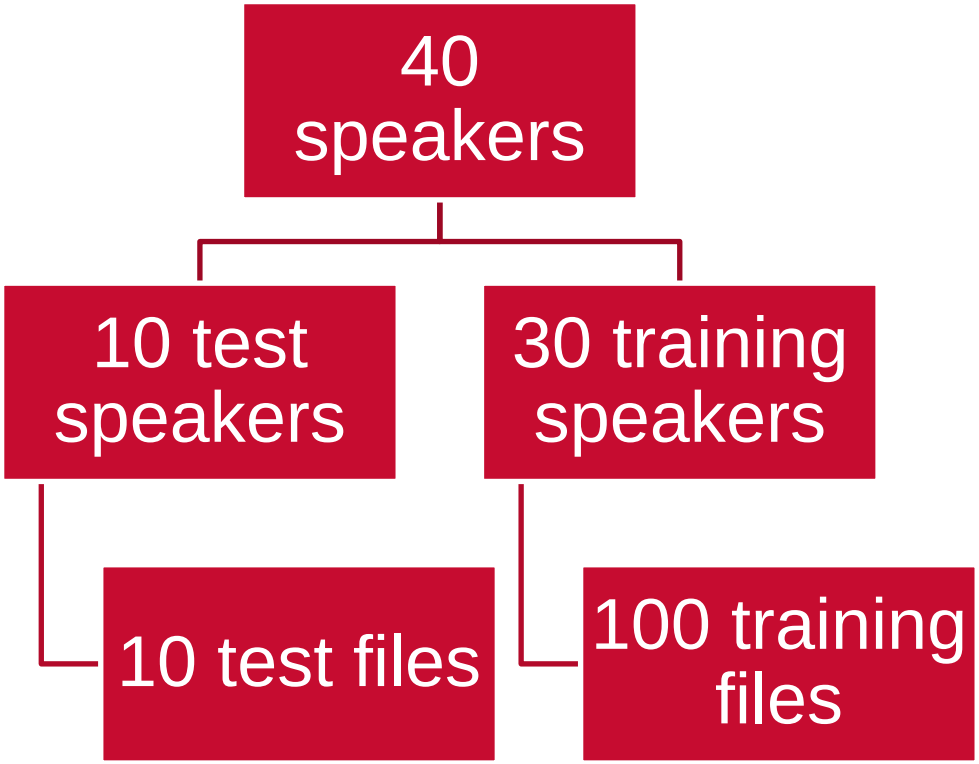
- 20 ms Hamming window, advanced by 20 ms
- Normalized for channel effects by dividing by overall RMS energy

LPC residual energy

- 25 ms window, advanced by 20 ms

Boakye, Kofi, et al. "Overlapped speech detection for improved speaker diarization in multiparty meetings." *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008.

Results: Training and Testing



- To ensure Leave-One-Speaker-Out Cross-validation

Results: GMM-HMM on audiobooks

- Confusion matrix of prediction with 20 ms window

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	98.68 %	1.32 %	0.00 %
	Single-speaker	1.07 %	61.87 %	37.06 %
	Overlap	0.00 %	26.59 %	73.41 %

Results: GMM-HMM on audiobooks

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	97.12 %	2.76 %	0.12 %
	Single-speaker	1.47 %	65.05 %	33.48 %
	Overlap	0.00 %	32.47 %	67.53 %

Confusion matrix for 50 ms window

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	95.78 %	4.22 %	0.00 %
	Single-speaker	1.89 %	76.83%	21.28 %
	Overlap	0.00 %	38.94 %	61.06 %

Confusion matrix for 100 ms window

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	96.09 %	3.91 %	0.00 %
	Single-speaker	1.47 %	87.21 %	11.32 %
	Overlap	0.07 %	48.33 %	51.60 %

Confusion matrix for 150 ms window

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	96.70 %	3.30 %	0.00 %
	Single-speaker	1.37 %	90.42 %	8.21 %
	Overlap	0.00 %	57.89 %	42.11 %

Confusion matrix for 200 ms window

Results: CRF on audiobooks

- Conditional Random Fields – another algorithm to deal with time-series data

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	92.27 %	7.73 %	0.00 %
	Single-speaker	0.36 %	96.13 %	3.51 %
	Overlap	0.00 %	88.85%	11.15 %

Results: GMM-HMM on real data

- NVAC audio data collected during Sociofeedback experiments
- Originally dual-channel, converted to mono-channel
- Tested on 5 speakers in a LOO fashion

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	67.77 %	25.71 %	6.52 %
	Single-speaker	10.61 %	50.62 %	38.77 %
	Overlap	1.51 %	35.85 %	62.64 %

Results: Challenges

- Overlap data amount is still very small compared to single-speaker data
- Overlap data still quite similar to single-speaker data with current features
- Output data requires post-processing to filter unnecessary switching between states
- Leave-one-speaker-out technique restricts developing particular speaker models

Conclusion

- Detecting overlap is essential to understanding human social behavior
- Current speaker diarization techniques do not detect overlap reliably
- Accuracy improvement through
  - Additional discriminative features
  - Output post-processing
  - Training speaker models for a particular constant speaker
  - Possibly other machine-learning algorithms

Acknowledgements

- Assoc. Prof. Justin Dauwels for his thoughtful insights
- Yasir Tahir and Hang Yu for their generous help
- IMI and IGS at NTU for supporting this research


THANK YOU
for listening
-
ANY
QUESTIONS?

Results: GMM-HMM on audiobooks and real data

- Trained on real data, tested on audiobooks

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	99.57 %	0.43 %	0.00 %
	Single-speaker	19.83 %	70.27 %	9.90 %
	Overlap	11.96 %	77.71 %	10.33 %

Results: GMM-HMM on audiobooks and real data

- Trained on audiobooks, tested on real data

		Predicted		
		Silence	Single-speaker	Overlap
Actual	Silence	0.00 %	93.94 %	6.06 %
	Single-speaker	0.00 %	71.91 %	28.09 %
	Overlap	0.00 %	48.87 %	51.13 %