

Combining Systems for High-accuracy Recognition and Diarization of Meetings

Andreas Stolcke

Amazon Alexa Speech

Sunnyvale, California

Research done at



stolcke@icsi.berkeley.edu

Roadmap

- Meeting transcription with ad-hoc arrays of devices (“virtual arrays”)
 - System architecture
 - Front-end processing
 - Speech & speaker recognition
 - System combination
 - Results on Denmark meetings
 - Results on NIST RT eval meetings

- System combination for speaker diarization
 - Problem statement and metric
 - DOVER algorithm
 - Results on multi-microphone meeting recordings
 - Results on single audio stream

Meeting Transcription with Ad-hoc Microphone Arrays

Collaborators: Takuya Yoshioka, Zhuo Chen, Dimitrios Dimitriadis,
William Hinthorn (Microsoft)

Meeting transcription: The challenge

Multiple (> 2) speakers

Unconstrained, conversational style

Face-to-face interaction, overlapping speech

Distant microphone capture / room acoustics

- Reverberation
- Background noise

The goal:

Meeting transcription from distant microphones as good as if each speaker were captured by a head-worn microphone

Approach 1 (1990-2000s)

Close-talking (head-worn) microphones

Not practical for business and consumer scenarios



Meeting recording at NIST

Approach 2

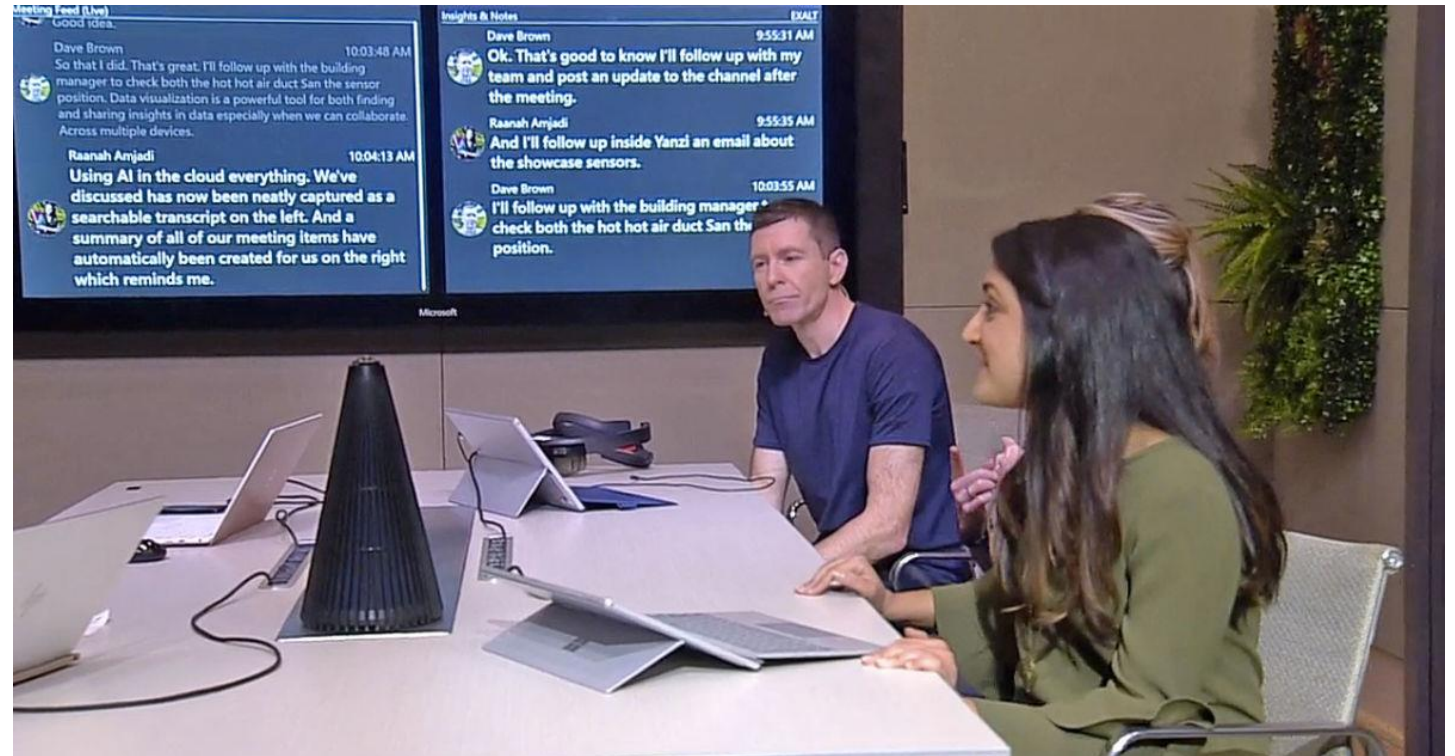
Microphone array device

Allows beamforming and noise cancellation for distant speech capture

Not suitable for:

- Consumer mass market
- Small budgets
- Ad hoc meetings

Princeton device demo,
Microsoft //build 2018



Meetings in the age of the smartphone ...



“Someone still has his cellphone on.”

Approach 3: Project Denmark

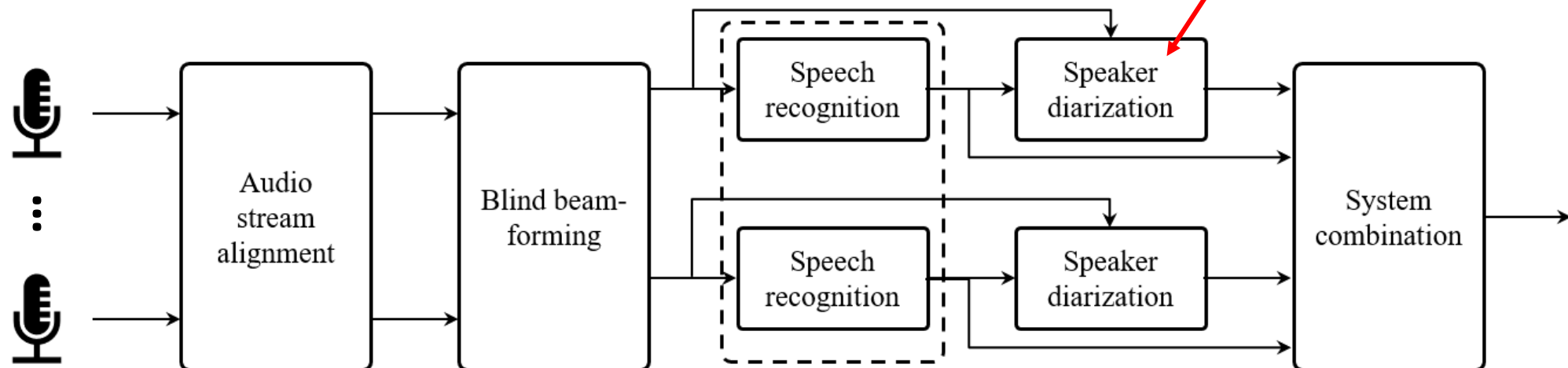
Multiple consumer-level mobile devices
Dynamically assemble ad-hoc microphone array
Audio synchronized and processed in the cloud

Yoshioka et al. , Interspeech 2019 and
Tech. Report MSR-TR-2019-11 [[arXiv:1905.02545](https://arxiv.org/abs/1905.02545)]

More information at

<https://www.microsoft.com/en-us/research/project/project-denmark/>

Based on enrolled
speaker ID:
Not discussed
here



Denmark Demo (from //build 2019)



Audio processing

Stream alignment

- Picks one audio stream as reference, aligns all others to it
- Every 30 seconds, estimate time lag between streams by maximizing cross-correlation
- More frequent sync in the beginning to establish global offset

Blind beamforming

- Combines multiple signals to enhance speech, attenuate noise
- Estimated by MVDR (minimum variance distortionless response)
- Reestimates beamformer coefficients every second, separately for different frequency bins

Interaction of beamforming and system combination

Beamforming makes resulting audio streams more correlated

Hypothesis combination is less effective if input hypotheses are decorrelated (errors are no longer independent)

Create multiple, diverse beamforming outputs by

- Rotating the reference channel among the various inputs
- Leaving out one input channel at a time (use $M-1$ out of M channels) [Stolcke, ICASSP 2011]

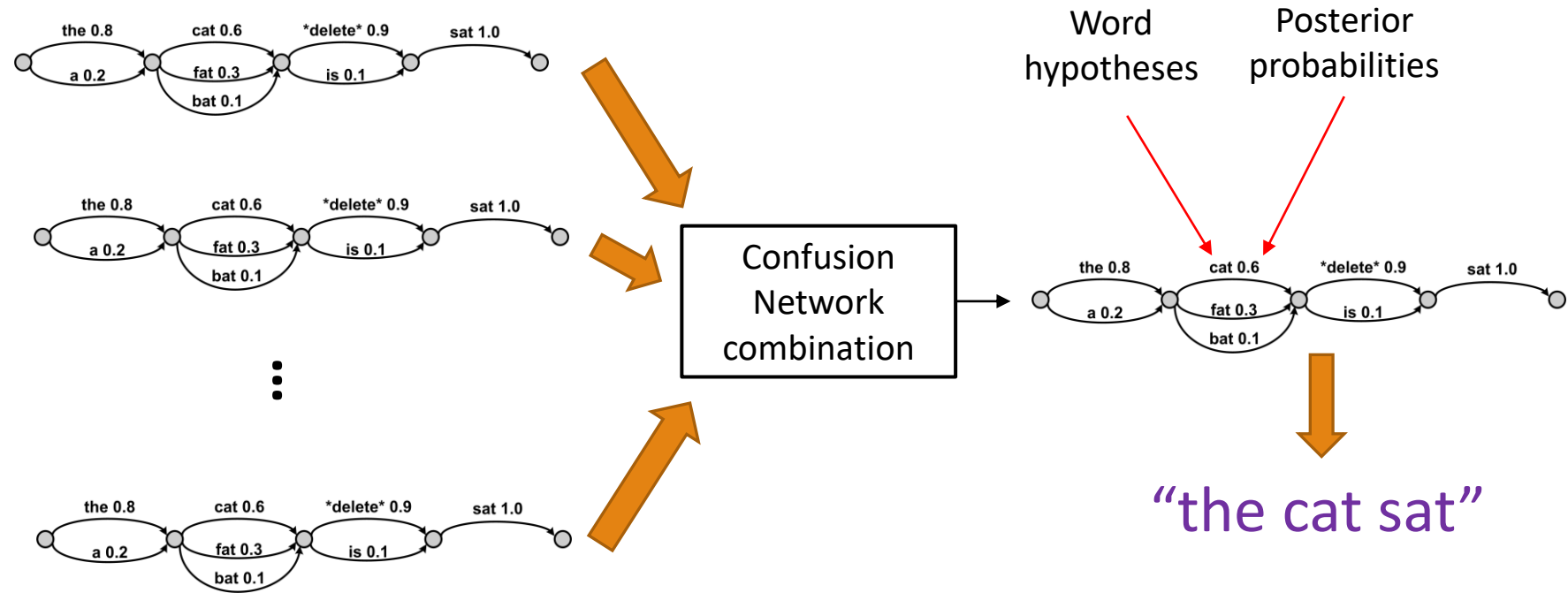
Leave-one-out beamforming requires inverting M different $(M-1)$ -dimensional spatial covariance matrices

- Avoid extra computation by deriving $(M-1)$ -dim inverse covariances from single M -dim inv. cov. matrix

Confusion Network Combination

Modifications to CNC:

- CNs are concatenated between segmentation points common to all streams [ICASSP 2010]
- Soft time mismatch penalty (in addition to edit distance cost)
- CNs encode words and speakers



Merging speaker and word recognition

Word reco 1:

unk-spkr 1.0	the 0.8	unk-spkr 1.0	cat 0.6	unk-spkr 1.0	*delete* 0.9
	a 0.2		fat 0.3		is 0.1
			bat 0.1		

Speaker reco 1:
(with 1-best words)

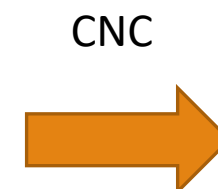
spkrA 0.7	the 0.01	spkrA 0.8	cat 0.01	spkrB 0.6	*delete* 0.01
spkr B 0.3		spkrB 0.2		spkrA 0.2	
				spkrC 0.1	

Word reco 2:

unk-spkr 1.0	the 0.5	unk-spkr 1.0	cab 0.4	unk-spkr 1.0	*delete* 0.4
	that 0.5		cat 0.3		is 0.3
			bat 0.2		It 0.3

Speaker reco 2:
(with 1-best words)

spkrA 0.5	the 0.01	spkrA 0.7	cab 0.01	spkrA 0.6	*delete* 0.01
spkrB 0.5		spkrB 0.3		spkrB 0.3	
				spkrC 0.1	




Result of CNC with speaker and word info

Confusion network alignment with:

- disallow aligning speaker labels to word labels
- disallow “unknown speaker”

CNC



spkrA 1.3	the 1.3	spkrA 1.5	cat 0.9	spkrB 1.0	*delete* 1.3
spkr B 0.8	A 0.7	spkrB 0.5	cab 0.4	spkrA 0.8	Is 0.4
			fat 0.3	spkrC 0.2	It 0.3
			bat 0.3		



1-best decoding

spkrA: the cat
spkrB: ...

Denmark meeting test set

5 unscripted work meetings, duration 0.5...1 hour each, 3...11 speakers

- 3 meetings: recorded with 4 different iOS devices, 3 different Android devices
- 2 meetings: processed the raw signals from Princeton microphone

7 devices/audio channels per meeting

10% of speech duration had more than one speaker overlapping

Speech recognition used pre-existing Microsoft conversational transcription service decoder and models

WER scored with NIST “asclite” tool (aligns single word hypothesis stream to multiple parallel reference transcripts)

We also evaluated speaker-attributed WER (SA-WER)

- A word must have the right speaker label to be counted as correct

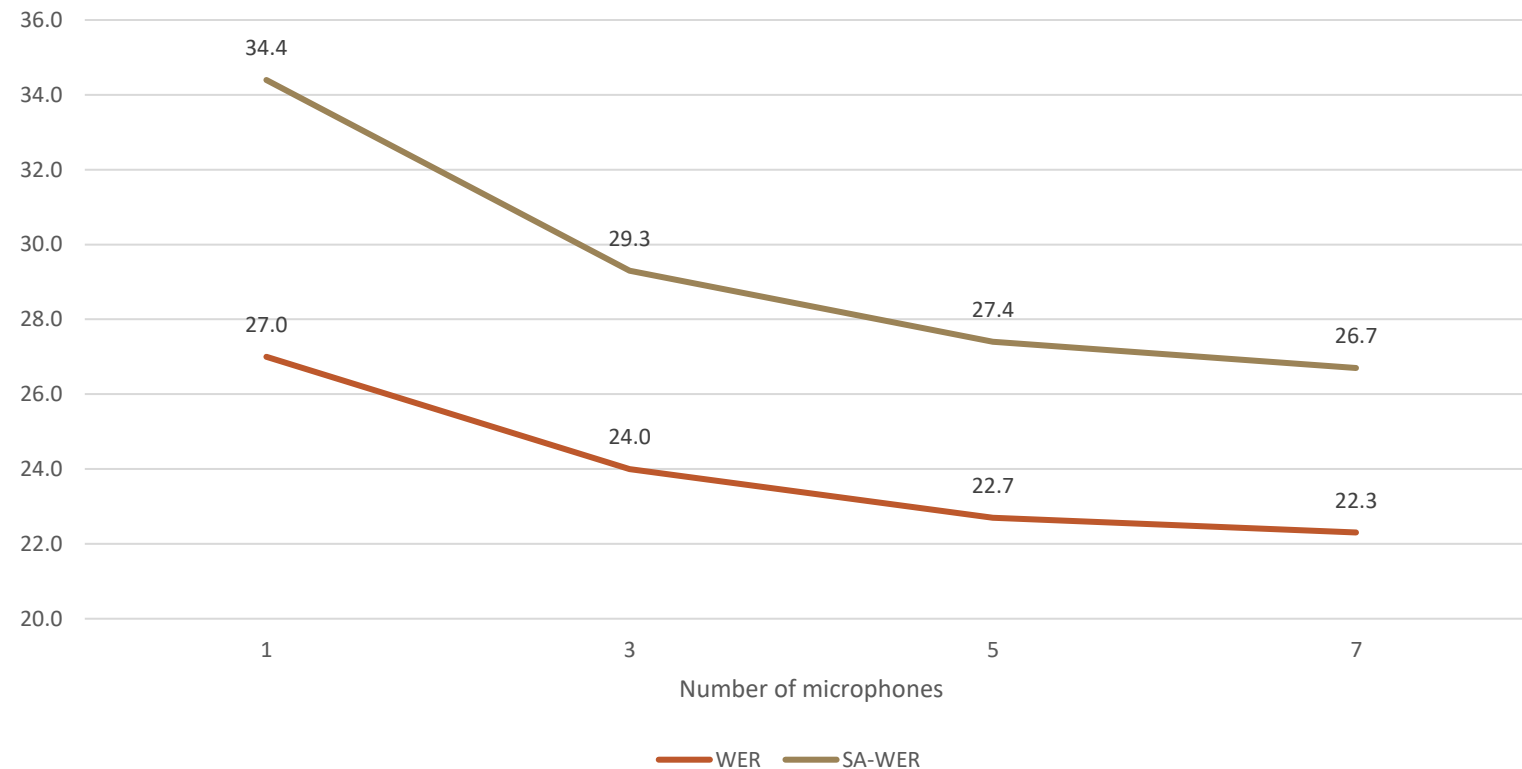
Results: Beamforming and CNC

Beamforming (7 microphones)	System Combination	WER	SA-WER
None	None*	27.0	34.4
All channels	None*	24.8	30.8
Leave-one-out	None*	24.9	30.9
None	CNC	22.8	27.7
All channels	CNC	22.5	26.9
Leave-one-out	CNC	22.3	26.7
<i>Close-talking microphones</i>	None	14.4	

* Average over all 7 channels

Importance of multi-channel processing

Results with leave-one-out beamforming and CNC



Closing in on close-talking

WER on **non-overlapping** speech segments

System / microphones	WER
1 microphone (average)	20.6
Beamforming, 7 microphones	18.1
Beamforming + CNC, 7 microphones	16.2
Close-talking microphones	13.2

} 3.0% absolute Δ

NIST 2007 Rich Transcription evaluation

8 conference meetings, 22-minute excerpts are transcribed and evaluated

4 different recordings sites

Number of microphones varies from 3 to 16

Input channels are already synchronized

- Denmark front-end was run unchanged
- Word duplicate removal (after system combination) was disabled

Three evaluation conditions:

- SDM: single distant microphone (“centrally located”)
- MDM: multiple distant microphones (allows beamforming, system combination, etc.)
- IHM: individual head-mounted microphones (close-talking)

“Meeting Transcription Using Virtual Microphone Arrays”, arxiv:1905.02545 (MSR Tech Report)

NIST RT-07 word error rates

Evaluation Condition	Overlap ≤ 4	No overlap
SDM	28.2	16.7
MDM with CNC	26.2	15.5
MDM with all-mic BF and CNC	26.3	14.8
MDM with LOO-BF and CNC	26.0	14.6
IHM	15.9	12.3

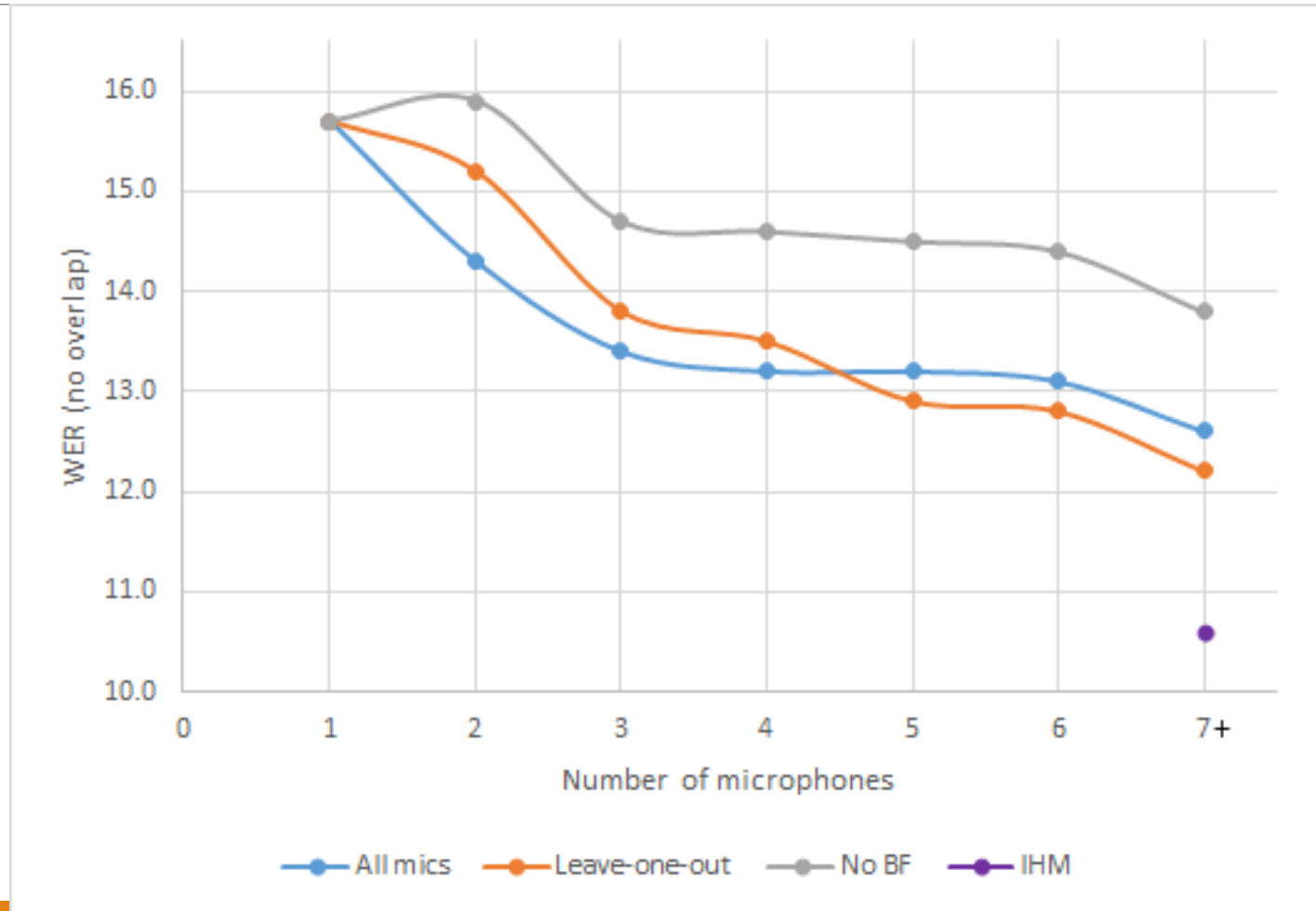
} 2.3% absolute Δ

SDM: single distant microphone (“centrally located”)

MDM: multiple distant microphones (allows beamforming, system combination, etc.)

IHM: individual head-mounted microphones (close-talking)

Effect of number of microphones



Summary:

Denmark meeting transcription

Multiple personal consumer devices can form microphone array that is

- Effective for capture and transcription
- Practical for a wide range of settings

Multiple levels of information fusion (front-end and hypothesis level) is key to success

Leave-one-out approach to beamforming works well system combination (CNC) if a sufficient number of input channels is available

- If not, beamform with all microphones

Our proof-of-concept system achieves accuracy within 2-3% absolute of close-talking recognition on non-overlapping speech

Overlapping speech (speech separation) is still a hard challenge

As is speaker diarization without prior enrollment!

DOVER:

System combination for diarization

Collaborator: Takuya Yoshioka

The speaker diarization task

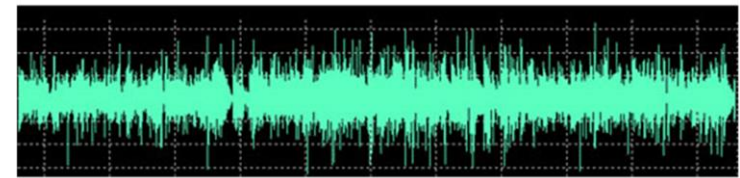
Task: “Who spoke When”

No prior knowledge of speakers

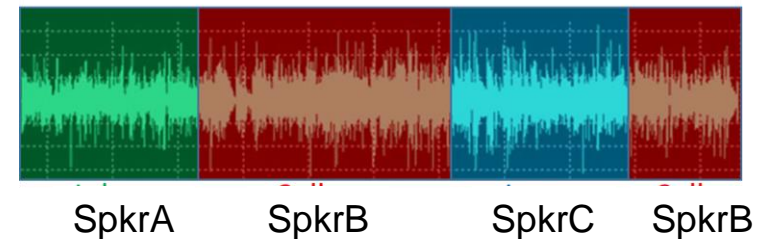
Important for:

- Interpreting the words (speaker-attributed transcripts)
- Understanding interaction among speakers
- Speaker adaptation

Input



Output



DER: Speaker diarization evaluation

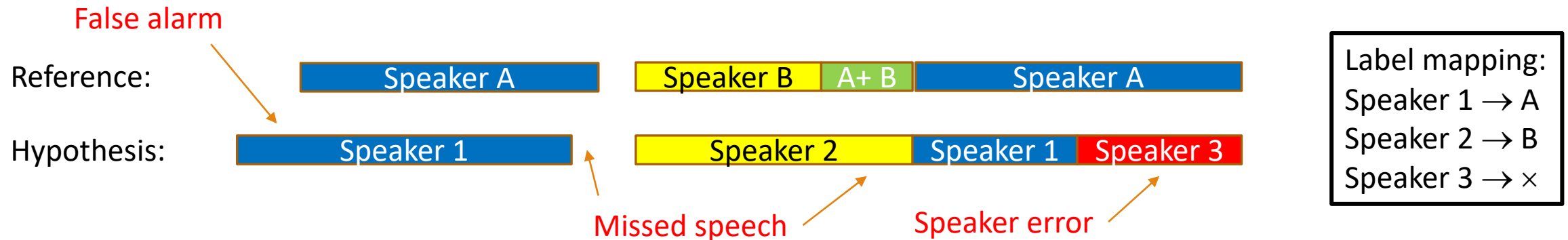
Diarization error rate defined as

$$\text{DER} = \frac{\text{Missed speech duration} + \text{False alarm speech duration} + \text{Speaker error duration}}{\text{True speech duration}}$$

Multiple overlapping speakers are scored individually

- To be perfect, system must recognize all overlapping speakers

Scoring tool finds an optimal mapping from hypothesized speaker labels to reference labels



Denmark diarization results

Denmark currently requires speaker enrollment (i.e., no true diarization)

We score speaker recognition output as diarization (speaker labels treated as anonymous)

System does not try to label overlapping speakers

- 10.0% of total speech duration = floor on missed speech and DER

Speech activity detection is performed by the transcription system

- Added 0.5s padding at the margins

	Missed speech	False alarm	Speaker error	DER
Avg. by channel	10.5	3.3	1.8	15.6
CNC output	10.2	2.4	1.0	13.6

Ensemble/voting methods

Multiple classifier are better than one

Combine output of different classifiers by

- Voting (majority wins)
- Score combination (soft voting), such as
- Posterior probability interpolation)
- Optionally, inputs can be weights

Widely used in speech recognition

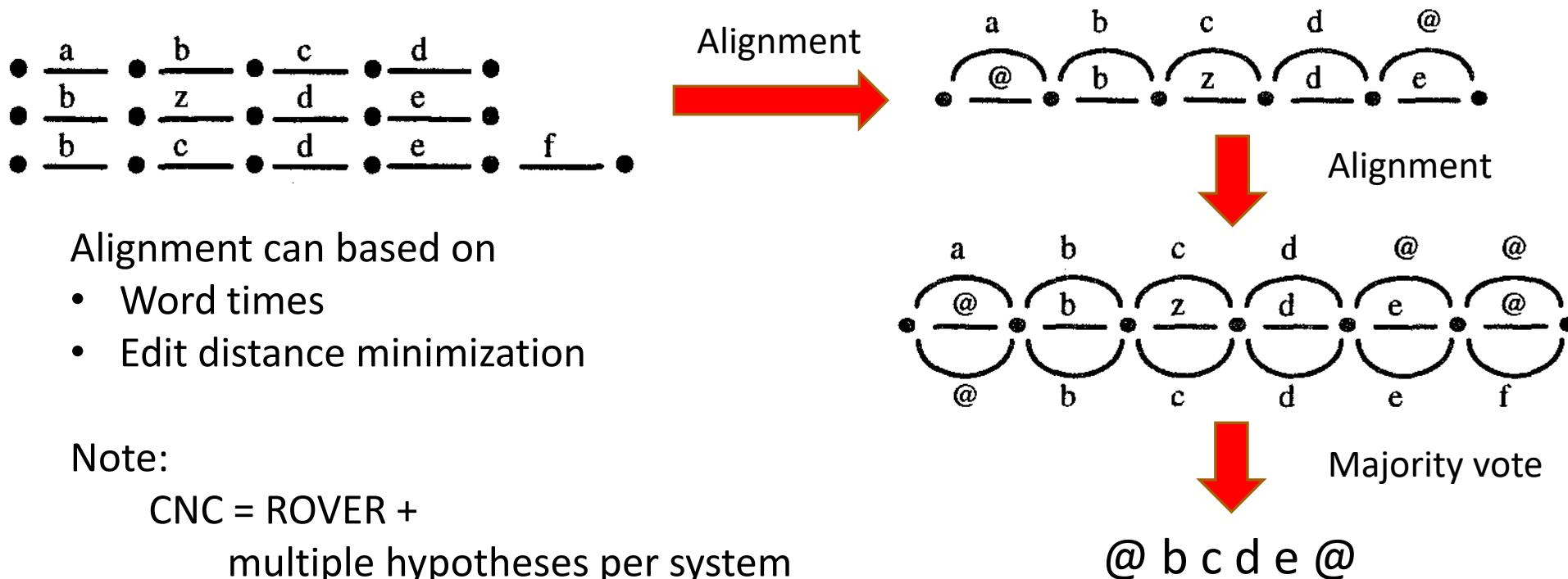
- ROVER
- Confusion network combination (CNC)

Almost always lowers the error when inputs are about equally good, but different/independent

A POST-PROCESSING SYSTEM TO YIELD REDUCED WORD ERROR RATES: RECOGNIZER OUTPUT VOTING ERROR REDUCTION (ROVER)

ASRU 1997

Jonathan G. Fiscus
National Institute of Standards and Technology
Gaithersburg, MD 20899



Alignment can be based on

- Word times
- Edit distance minimization

Note:

CNC = ROVER +
multiple hypotheses per system

Approach 1 (1990-2000s)

Close-talking (head-worn) microphones

Not practical for business and consumer scenarios



Jon Fiscus

Meeting recording at NIST

DOVER:

Diarization Output Voting Error Reduction

How can we vote among a set of diarization outputs?

Problem: output labels are unrelated between different systems

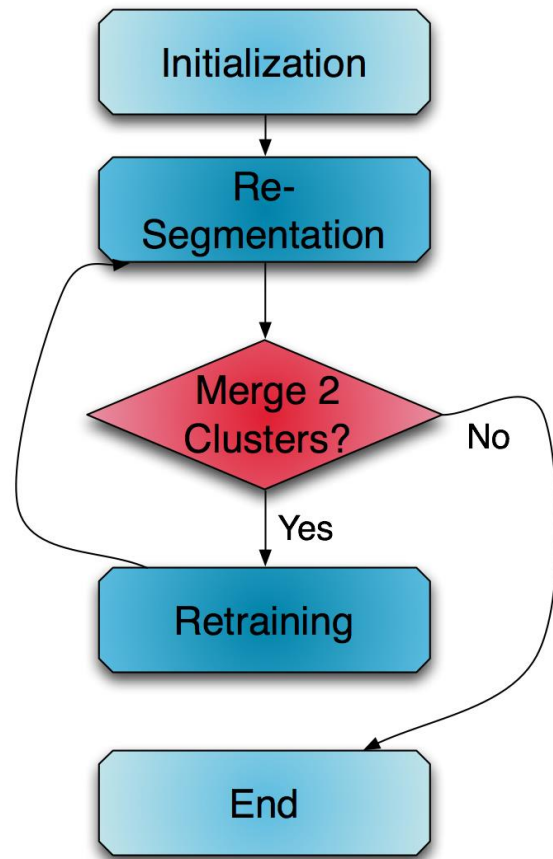
This is the same problem as for scoring diarization output against a reference

Solution: perform minimum cost mapping of labels into a common label vocabulary

1. Initial alignment = First diarization output
2. While there are more diarization output:
 - a. Map next output to existing alignment labels, minimizing DER
3. For all time instances, output majority label

“DOVER: A Method for Combining Diarization Outputs”, arXiv:1909.08090 (ASRU-2019)

Speaker clustering algorithm (IDIAP/ICSI)



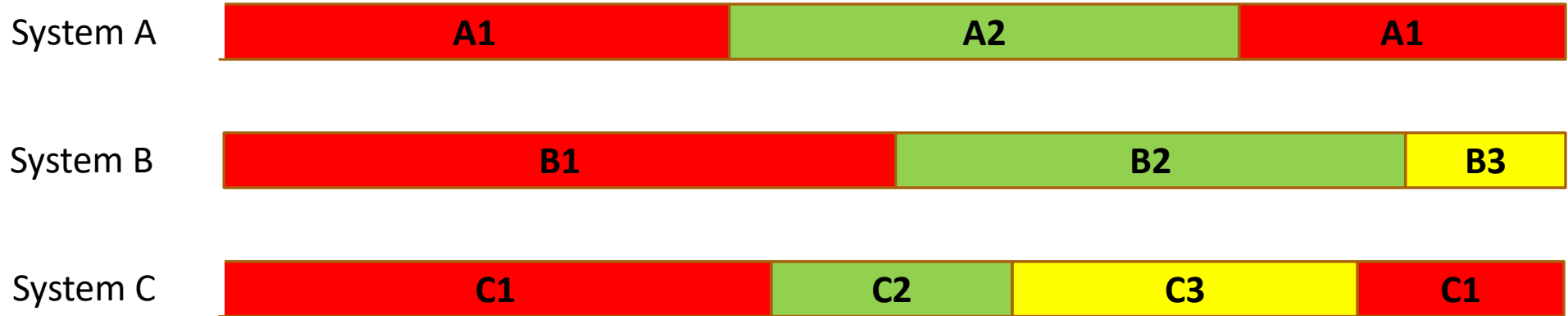
1. Create k random segments and train k GMMs with g Gaussians each
2. Assign frames to clusters according to likelihoods
3. Use Bayes information criterion (BIC) to determine if two clusters should be merged

Ajmera, McCowan & Bourlard, "BIC revisited for speaker change detection", IDIAP, 2002

Wooters & Huijbregts, "The ICSI RT07s speaker diarization system", MLMI 2007

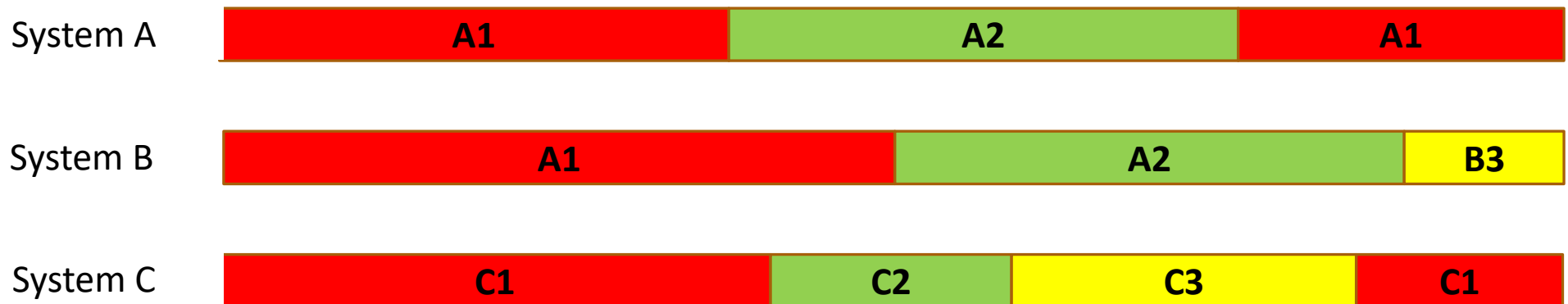
An example

1 - Inputs = original labels



2 - Map System B to System A labels

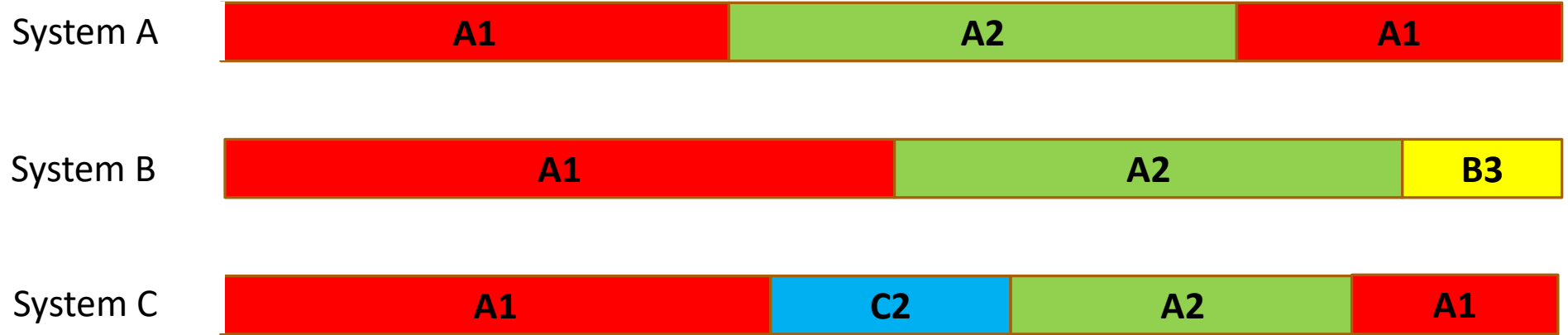
- B1 → A1
- B2 → A2
- B3 not mapped



Example, continued

3 - Map System C labels to System A+B labels

- C1 → A1
- C2 not mapped
- C3 → A2



4 - Voting



If inputs differ on speech activity, output speech if $\geq \frac{1}{2}$ have speech

Tie (no majority)

Anchoring and tie-breaking

Label mapping is greedy, dependent on ordering of inputs

As with word string alignment, best to start with the highest-accuracy hypothesis (anchor)

Heuristic: start with the *centroid* (shortest distance to all other hypotheses)

1. Compute average DER between each hyp and all others
2. Rank hyps by average DER, least first
3. Weight each hypothesis by $\left(\frac{1}{\text{rank}}\right)^{0.1}$
4. Apply DOVER (with weighted voting step)

Rank-based weighting breaks ties in favor of more reliable inputs

- but two lower-ranked hypotheses can still overrule a single higher-ranked hypothesis

Duality of DOVER and ROVER

	DOVER	ROVER
<i>Alignment of</i>	speaker labels	word labels
<i>sharing a common</i>	time axis	label space (vocabulary)
<i>in</i>	label space	time
<i>minimizing cost of</i>	diarization error	word error (string edit distance)
<i>followed by voting within each</i>	speaker segment	word confusion set

Experiments

Two meeting datasets: NIST RT-07 and Denmark (same as for transcription experiments)

Speech activity detection same for all inputs

- NIST RT-07: SRI/ICSI 2007 eval system SAD
- Denmark: padded ASR output

Variety of features streams:

- MFCC from raw audio (19-d, every 10 ms)
- MFCC from beamformed audio (using leave-one-out)
- Time delay of arrival features (TDOA, estimated by BeamformIt tool) [Anguera et al. 2007]
- D-vectors computed by Denmark for speaker ID (128-d mapped to first 30 principal components)

Speaker clustering using IDIAP/ICSI agglomerative algorithm

Process all audio streams independently, then DOVER

DER results on NIST RT-07

Diarization features	DOVER inputs			DOVER outputs	
	Max	Average	Min	SpkrErr	DER
MFCC (raw audio)	21.69	14.13	8.41	10.39	18.91
MFCC (beamformed audio)	16.80	9.43	5.48	7.04	15.58
MFCC + TDOA	12.79	5.30	2.16	2.38	10.93

Max/average/min are over all input channels (min = oracle choice)

Missed speech rate = 3.9%

False alarm rate = 4.6%

DER results on Denmark meetings

Diarization features	DOVER inputs			DOVER outputs	
	Max	Average	Min	SpkrErr	DER
MFCC (beamformed audio)	34.56	23.23	15.56	15.00	26.94
MFCC + d-vector	13.94	11.06	8.82	8.70	20.65
MFCC + 3 d-vectors*	11.38	6.07	3.00	3.10	14.97
Speaker ID (using enrolled speakers)	2.18	1.86	1.42	1.20	13.06

Missed speech rate = 11.3% (overlapped speech: 10.0%)

False alarm rate = 0.6%

* Channel i used d-vectors from channels $i - 1, i, i + 1 \pmod{7}$

DOVER for Single Audio Stream

Create multiple diarization hypotheses from a single input by

- Varying hyperparameters of the clustering algorithm, for example
 - Number of initial clusters
 - interpolation weight for feature streams (MFCC vs. TDOA)
- Introducing pseudo-randomness (flipping coins)

Then use DOVER to combine the outputs

Data used: beamformed distant microphones from NIST RT meeting evals

- Tuning set: RT-07
- Eval set: RT-09

“Improving Diarization Robustness using Diversification, Randomization and the DOVER Algorithm”,
arXiv:1910.11691

DOVER on Randomize Clustering

Modified the best-first speaker clustering algorithm

With probability 0.3, pick the second-best pair of clusters to merge at each iteration

Method	Seed	RT-07 SpkrErr	RT-09 SpkrErr
Best first		4.1	8.5
Randomized	1	2.8	7.5
	2	2.4	8.1
	3	3.7	8.3
	4	3.6	8.7
	5	5.4	8.5
DOVER	1+2+3+4+5	3.3	8.1

DOVER with varying TDOA stream weight

TDOA weight	RT-07 SpkrErr	RT-09 SpkrErr
0.715	2.9	7.7
0.720	2.6	7.5
0.725	2.8	9.0
0.730	5.7	7.9
0.735	5.7	7.2
0.740	3.7	7.9
0.745	2.6	7.7
0.750	4.1	8.5
0.755	2.8	7.5
0.760	2.8	7.5
DOVER	2.5	7.4

DOVER with varying initial cluster no.

Initial no. clusters	RT-07 SpkrErr	RT-09 SpkrEr
16 (default)	4.1	8.5
18	2.6	7.4
20	2.5	7.4
22	3.1	7.2
24	5.5	6.7
DOVER	2.1	6.5

Summary:

Combining diarization outputs

DOVER: new algorithm for weighted voting among diarization systems

Dual to ROVER: align hypotheses in label space instead of time, then vote

Ideal for combining outputs from independent diarization of multiple audio channels

Results on NIST RT-07 and Denmark meetings:

- Diarization is highly sensitive to choice of channels (even after beamforming)
- DOVER output close to, or better than, oracle-choice channel
- Consistent for a variety of input features
- Even improves on output of speaker ID

Challenges:

- Hybrid diarization of enrolled and unknown speakers
- Overlapped speech

Summary (continued)

On single audio, DOVER can help by combining multiple diarization hypotheses

- Clustering is very sensitive to hyperparameters, and tuning does not generalize across data sets
- DOVER of multiple runs give robustness

Randomizing the clustering and combining with DOVER yields higher accuracy than best-first clustering

Questions?
