Coupling Beamforming with Spatial and Spectral Feature Based Spectral Enhancement and Its Application to Meeting Recognition

Tomohiro Nakatani, Mehrez Souden, Shoko Araki, Takuya Yoshioka, Takaaki Hori, Atsunori Ogawa

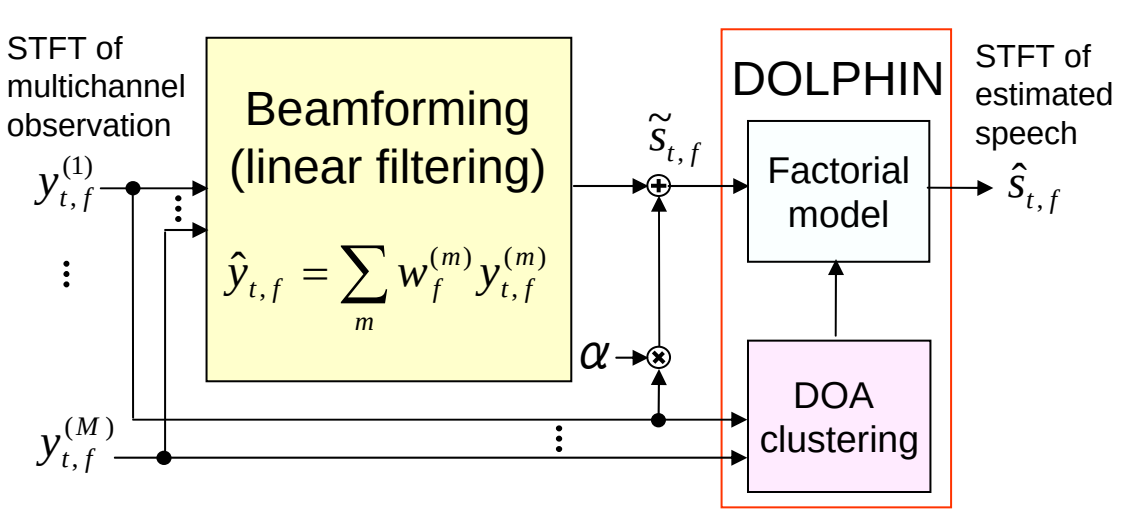
OVERVIEW

Goal:
Speech enhancement in highly nonstationary interference environment

ICASSP-2012
DOLPHIN --- spectral and direction-of-arrival (DOA) feature based enhancement

ICASSP-2013 (This paper)
Coupling beamforming with DOLPHIN → Application to meeting recognition

PROPOSED FRAMEWORK: COUPLING BEAMFORMING with DOLPHIN



	Use of spatial characteristics	Use of spectral characteristics
Beamforming	☺ (Fully utilized by linear filtering)	☹ (No a priori knowledge)
DOLPHIN	😐 (Partially utilized)	☺
Proposed coupling	☺	☺

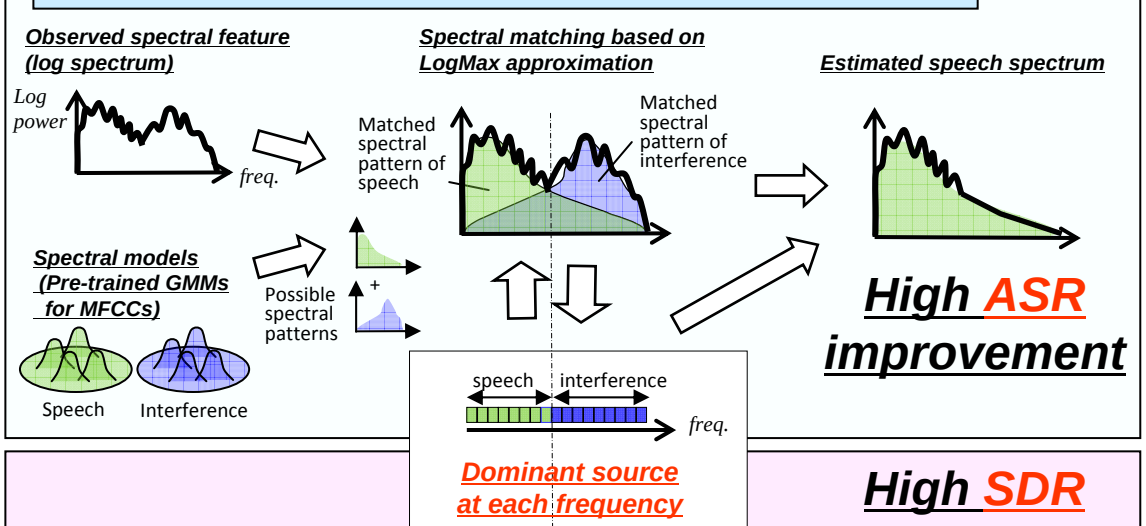
DOLPHIN*1 --- SPATIAL and SPECTRAL FEATURE BASED SPECTRAL ENHANCEMENT

*1) DOminance based Locational and Power-spectral cHaracteristics INtegration

- What is DOLPHIN
  Speech enhancement method that uses spatial and spectral characteristics of signals by integrating:
  - ***Factorial model based enhancement***
  - ***DOA clustering based enhancement***
- Benefit of integration
  Spectral matching by factorial model approach can be more reliable with DOA clustering
- Effectiveness
  Significant improvement of ASR and SDR*2
  → **Best keyword recognition score for 1st CHiME Challenge [Delcroix, 2013]**

  *2) Signal-to-Distortion Ratio
- Limitation
  It does not modify signal phase for enhancement
  → **Not yet fully utilize spatial characteristics of signals, e.g., for cancellation of point sources**

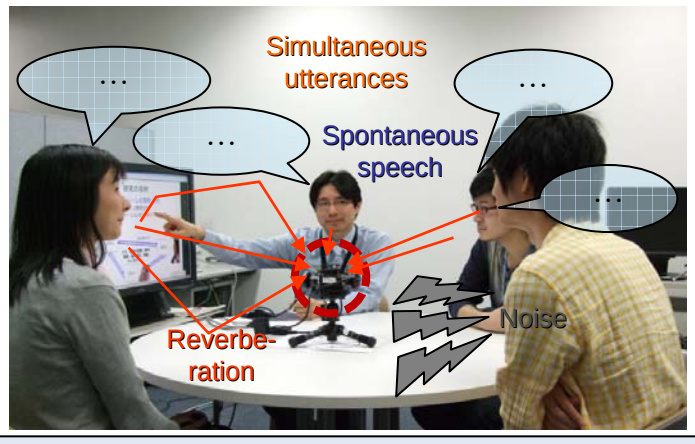
Factorial model based enhancement [Roweis, 2003]



DOA Clustering based enhancement [Yilmatz, 2004]

NTT

MEETING DATABASE [Hori, 2012]

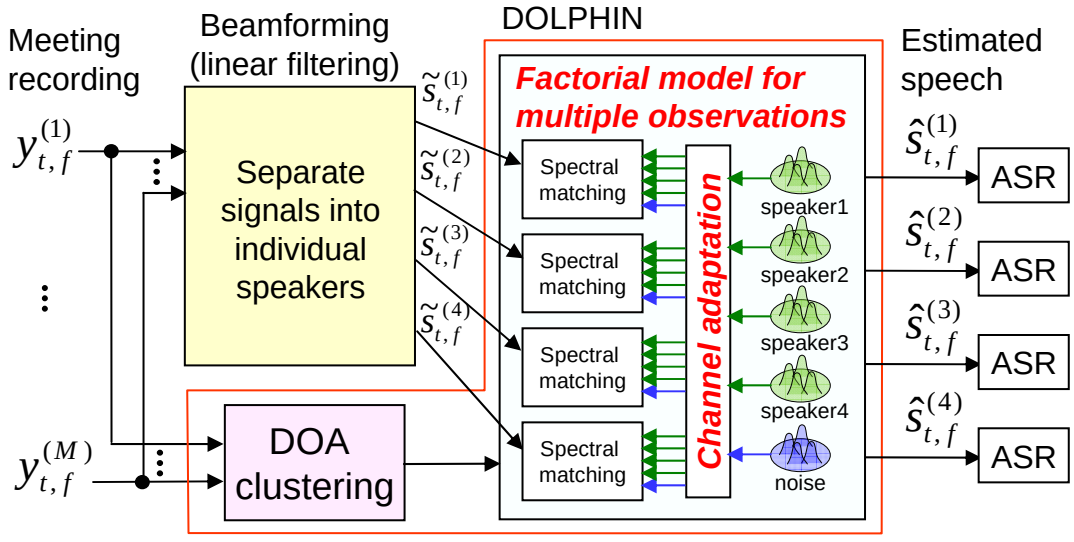


- Formal **spontaneous conversation**
  e.g. about "stress at your office"
- Recorded by **table microphones**
- Japanese
- Length of a session: about 15 min
- # of sessions (test set) : 16
- **Stationary Noise**
- **A lot of simultaneous utterances**

- Recording condition

	Office room	Sound-proof room
SNR	15 to 20 dB	20 to 25 dB
T60	350 ms	120 ms
# of mics	**8**	
# of speakers	4	
Mic-spkr dist.	1 m	

COUPLING for MEETING RECOGNITION



- Beamforming
  Blind source separation by linear filtering
  →All beamformer outputs contain all speakers and noise ***with modified acoustic channels***
- DOLPHIN
  Reduction of remaining interfering speakers and noise using DOA clustering and ***factorial model for multiple observations***
- ***Factorial model for multiple observations***
  Joint estimation of channels and spectral patterns to match all the beamformer outputs

MEETING RECOGNITION EXPERIMENTS

- Front-ends to be compared
  - ICA [FastICA+Natural Grad.]
  - MVDR [Souden, 2012]
  - DOLPHIN [Nakatani, 2012]
  - **ICA+DOLPHIN** (Proposed)
  - **MVDR+DOLPHIN** (Proposed)
- DOLPHIN training data
  - Corpus of Spontaneous Japanese (CSJ) recorded by a headset for speaker independent spectral model
- DOLPHIN channel adaptation
  - Unsupervised bias adaptation (batch)
- ASR training data
  - CSJ for acoustic model (trained based on dMMI criterion [McDermott, 2010])
  - CSJ, training set of meeting database (44 sessions), and WWW for language model (Total vocabulary size: 156k)
- ASR model adaptation
  - Unsupervised MLLR for acoustic model adaptation (batch)

Word error rates (%) obtained using different front-ends ***in office room / soundproof room*** w/ and w/o unsupervised MLLR for ASR

	w/o MLLR	w/ MLLR
Baseline (no front-end)	86.5 / 79.8	80.5 / 79.0
ICA	60.6 / 59.6	49.5 / 48.1
MVDR	47.1 / 52.6	37.6 / 43.9
DOLPHIN	49.2 / 48.9	41.1 / 45.1
ICA+DOLPHIN	43.8 / **45.6**	37.9 / **41.6**
MVDR+DOLPHIN	**40.6** / 48.0	**35.5** / 42.7
Headset (for reference)	30.6 / 40.7	27.0 / 36.1

CONCLUDING REMARKS

- Coupling of beamforming with DOLPHIN was shown to be very effective as a front-end for a large vocabulary meeting recognition task.
  - Beamforming: optimally controls linear filtering for enhancement based on spatial characteristics of signals
  - DOLPHIN: improves speech spectral estimates based on pre-trained spectral models and DOA clustering.

SOUND DEMONSTRATION

	Obs.	Estimated speech			
ICA					
MVDR					
DOLPHIN					
ICA+DOLPHIN					
MVDR+DOLPHIN					

Check them on our laptop