MULTI-GEOMETRY SPATIAL ACOUSTIC MODELING FOR DISTANT SPEECH RECOGNITION

Kenichi Kumatani, Minhua Wu, Shiva Sundaram, Nikko Ström, Björn Hoffmeister

Acknowledgements: Arindam Mandal, Brian King, Chris Beauchene, Gautam Tiwari, I-Fan Chen, Jeremie Lecomte, Lucas Seibert, Roland Maas, Sergey Didenko, Zaid Ahmed

Abstract

Goal:

- Building a single acoustic model that can cover multiple array geometries
- Making the model optimal for far-field automatic speech recognition (ASR)
- Achieving real-time processing without any non-causal processing pass

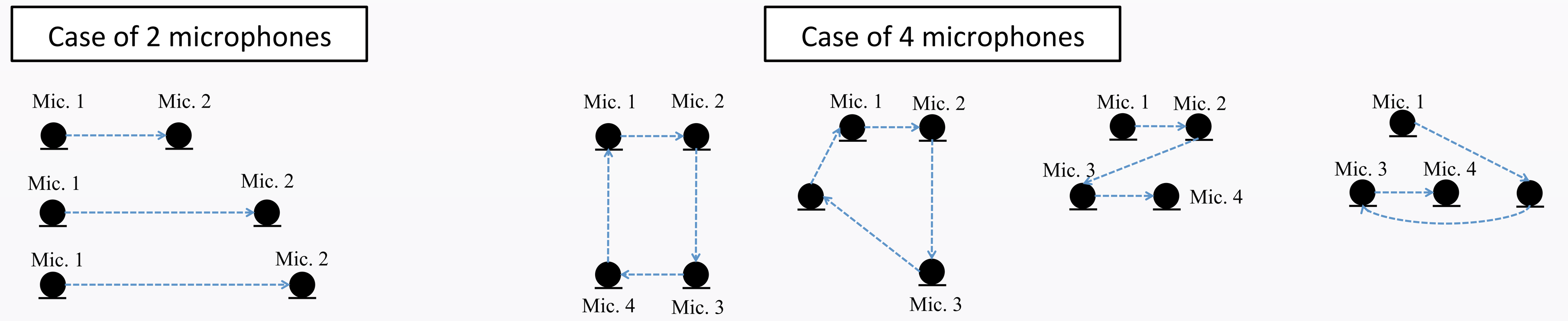
Our approach:

- Extending our work [1] (being presented in the same session) so as to model multiple array geometries and
- Training the multi-geometry array front-end and phone classifier jointly with the real-world data.

Technical Issue

What is array geometry?

The array geometry can be specified with relative positions to a reference microphone. each figure below shows a different array geometry structure.



Array geometry mismatch

The noise suppression and speech enhancement performance degrades when there is a mismatch between training and test array geometry conditions.

Conventional solutions

Method	Need Supervised Signal?	Need Adaptation Data?	Possible disadvantage and citation
Self-calibration	Yes	No	Supervised signal such as swept frequency signal will need to be played and captured [2]
Calibration with noise field	No	Yes	A noise field must be assumed [3]
Microphone selection	No	Yes	Ignoring sensors will compromise the best possible noise suppression performance [4]
Feature-based approach	No	No	This will not maximize the benefit of multi-channel information [5]
Blind estimation	No	Yes	One utterance data or more is required for maintaining performance [6]
Multi-style training	No	No	Multiple array geometry information is not usually incorporated into the network [7]

See also the papers for more details.

Our strategy

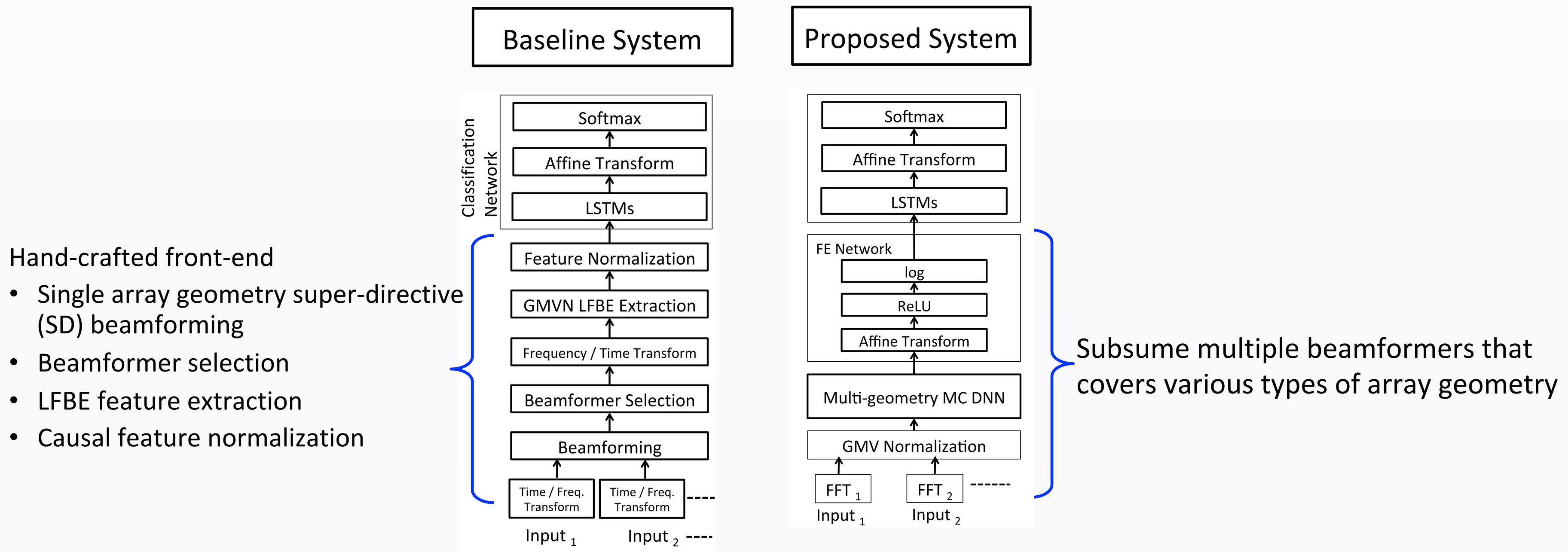
- We architect the neural network so as to model multiple array geometry structures; the multi-geometry network will be trained with multi-geometry array data so as to maximize the phone classification error.
- Everything will be learned from the real far-field data so that it neither requires supervised signal or adaptation data.
- In contrast to conventional multi-style training, it will embedded the sound propagation model into the network.

References

[1] Wu Minhua, Kenichi Kumatani, Shiva Sundaram, Nikko Ström, Björn Hoffmeister, "Frequency Domain Multi-channel Acoustic Modeling for Distant Speech Recognition", ICASSP 2019.
[2] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, Chichester, UK, 2009.
[3] I. McCowan et al., "Microphone array shape calibration in diffuse noise fields," *IEEE Trans. ASLP*, 2008.
[4] K. Kumatani et al., "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Proc. HSCMA*, 2011.
[5] S. Braun et al., "Multi- channel attention for end-to-end speech recognition," in *Proc. Interspeech*, 2018.
[6] T. Higuchi et al., "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *Proc. ICASSP*, 2018.
[7] T.N.Sainath et al., "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *Proc. ASRU*, 2015.

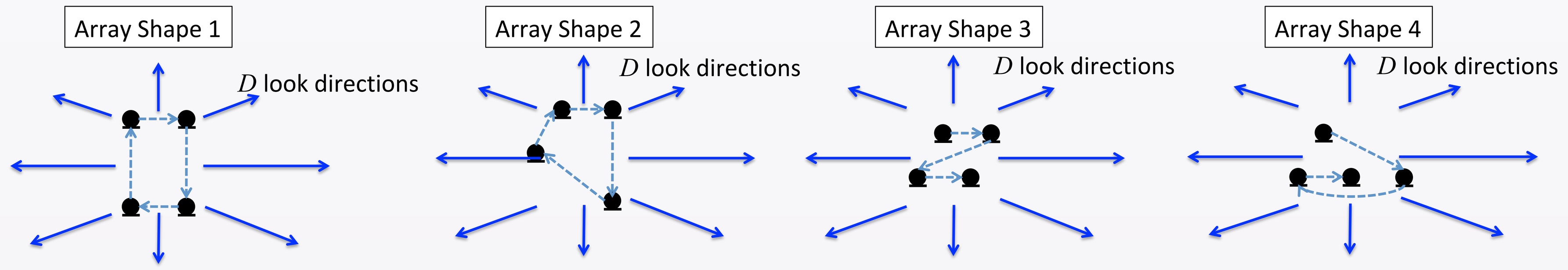
Our Far-field ASR system

- We initialize the multi-channel input layer with beamformers' weights calculated with different array configurations.
- The multi-geometry front-end is cascaded to and phone classifier without any speech reconstruction layer.
- The whole network is jointly optimized with array data of multiple geometries so as to minimize the phone discrimination error.



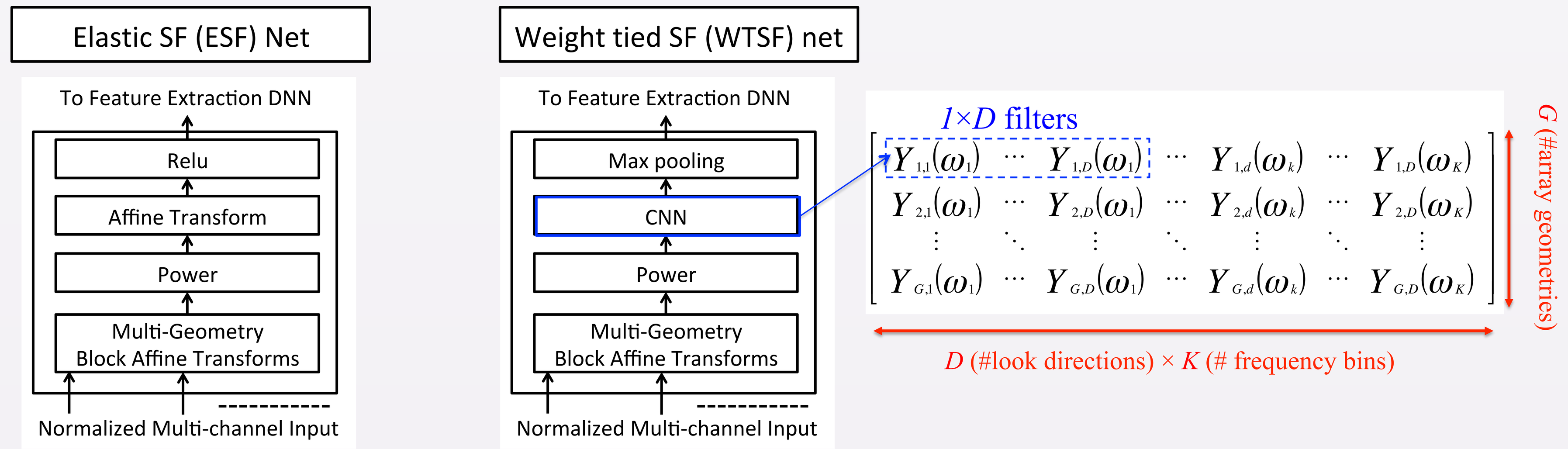
- Our whole multi-channel network is trained in a stage-wise manner; the classification layers are first trained with the log-filter-bank energy features (LFBE). The feature extraction and classification layers are then trained jointly with single channel DFT features. After we add spatial filtering layers initialized with super-directive (SD) beamformers' weights, we fine-tune the whole network with multi-channel DFT features of multiple array configurations.

Visualization of multi-geometry beamforming



The figures illustrates how beams are steered for a type of array geometry. In the case that we have G array shapes and build D beamformers for each, the total number of beamformers will be $D \times G$. We will need to efficiently combine or select them.

Multi-geometry spatial filtering (SF) network

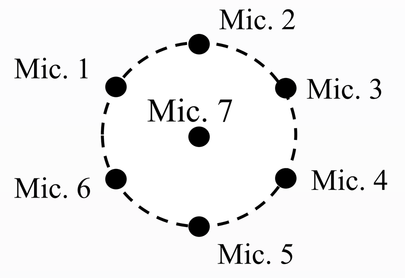


- We consider two network architectures: elastic SF (ESF) and weight-tied SF (WTSF) networks.
- The difference between two architectures is how spatial filtering layer output is combined;
  - ✓ The ESF network combines all the array output in a unconstrained weighted manner.
  - ✓ The WTSF net applies the same weight to all the frequency bins and picks the array output with the maximum energy.
- Notice that the WTSF net can reduce the number of parameters significantly.

ASR Experiments

- We used approximately 1100 hours of speech spoken by human beings, collected with the 7 microphone circular array in various rooms and split 1,000 and 100 hours into training and test sets where there is no overlapping speaker between sets
- Part of data are captured through a Live traffic where the interactions between the user and devices were completely unconstrained;
  - ✓ Users may move while speaking to the device.
  - ✓ Talker's position may change after each utterance.
- We observed that real-time adaptive beamforming degraded recognition accuracy due to steering errors [1]; we omit results of adaptive beamforming.

Change of array geometry



- We created different array geometry by selecting 2 or 4 sensors from 7 microphones.
- Two microphone case: clustering a pair of microphones based on microphone spacing
- Four microphone cases: grouping a set of congruent quadrilaterals and disordering the channels

Robustness against unseen array geometry

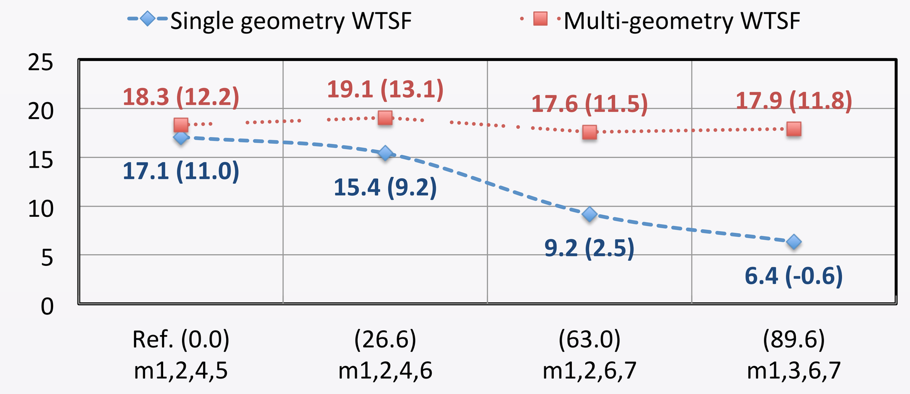
Modeling method	No. channels	No. mismatched sensor locations	WERR (%)		
			SNR>15	5 ≤ SNR < 15	SNR≤5
LFBE with single mic.	1	0	–	-	–
LFBE with SD BF	7	0	8.2 (–)	7.8 (–)	4.9 (–)
ESF with single geometry data	2	0	12.3 (4.5)	16.5 (9.5)	11.1 (6.6)
	2	1	10.0 (2.0)	15.0 (7.8)	9.8 (5.2)
ESF with single geometry data	4	0	16.4 (9.0)	21.7 (15.1)	15.5 (11.2)
	4	1	13.7 (6.0)	20.9 (14.3)	15.2 (10.9)
	4	2	6.8 (-1.5)	12.4 (5.0)	9.4 (4.8)
ESF with multi-geometry data: 2 sets of microphone spacing	2	0	11.6 (3.7)	16.7 (9.7)	11.4 (6.9)
	2	1	10.3 (2.2)	16.0 (9.0)	11.0 (6.5)
WTSF with multi-geometry data: 2 sets of microphone spacing	2	0	12.1 (4.2)	17.1 (10.1)	12.3 (7.8)
	2	1	11.0 (3.0)	16.0 (9.0)	11.8 (7.2)

- The recognition accuracy largely degrades in the mismatched geometry condition when the single geometry data are only used for training.
- Multi-geometry model can still maintain good accuracy in the mismatched geometry condition.
- The WTSF architecture achieve the best accuracy with a much less number of parameters than the fully-connected ESF network.

Coverage of different 4-channel array configuration

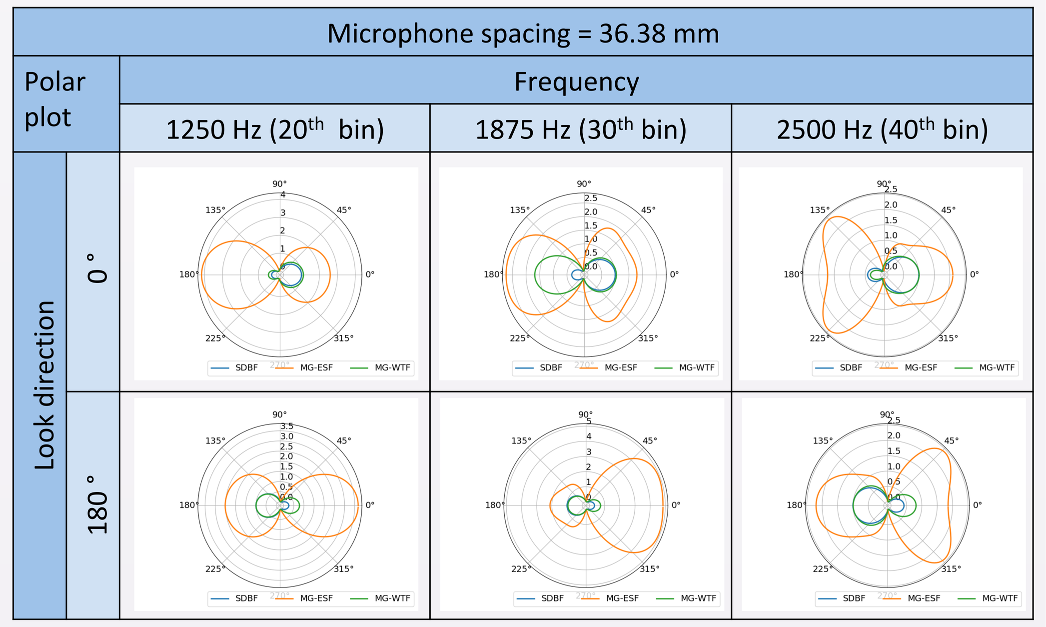
- There is significant degradation in the mismatched array configuration condition in the case of the single array geometry model.
- The degradation can be avoided by training the multi-geometry model.

The number in () indicates a dissimilarity index between two arrays which can be expressed as $\sum_{s=2}^{S} | d_{s,r}^{(1)} - d_{s,r}^{(2)} |$ where $d_{s,r}^{(i)}$ is the distance between the s^{th} and the reference sensors of the i[th] array.



Steering response power (SRP) w.r.t. a direction of arrival

- The left figure shows the SRP of SD beamforming (SD-BF), multi-geometry ESF (MG-ESF) and multi-geometry WTSF net (MG-WTF) for two-channel input.
- Each line indicates the directivity of the spatial filter, how much the filter strengthens or attenuates a signal coming from a particular direction.
- Notice that the ESF network will combine the spatial filters with weights in a soft-decision manner so as to maximize the phone classification accuracy; it may permute a look direction among different frequencies. it also tends to amplify the signal.
- The WTF network can avoid such a look direction inconsistency problem although it did not lead to recognition accuracy improvement.



Conclusion

- The fully-learnable multi-channel AM can learn multiple types of microphone array geometry.
- The multi-channel neural network trained with multi-array data can alleviate the mismatch between different array shapes.
- The model is also optimal in terms of speech recognition.
- The method neither requires adaptation process nor any bi-directional processing pass.