FREQUENCY DOMAIN MULTI-CHANNEL ACOUSTIC MODELING FOR DISTANT SPEECH RECOGNITION

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

amazon echo

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 an *optimal* acoustic model for far-field automatic speech recognition (ASR):
  - ✓ Achieving the better recognition accuracy with a fewer microphones,
  - ✓ Real-time processing without bi-directional processing or batch processing, and
  - ✓ Making a whole front-end learnable from a large amount of real-world data without risky adaptation.

Our approach:

- Unifying acoustic signal processing and ASR acoustic model with a fully learnable neural network
- Incorporating the sound propagation model into a neural network

Background

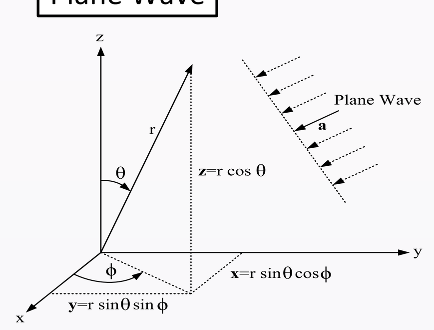
Why do we need multiple microphones for far-field speech recognition?

We can leverage spatial information by measuring sound pressure at multiple points, which enables us to

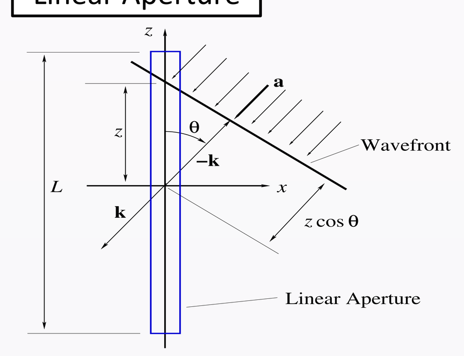
- Suppress interfering signals based on a direction of signal arrival, and
- Maintain the minimum distortion amount for a look direction

Far-field wave propagation model

Plane Wave

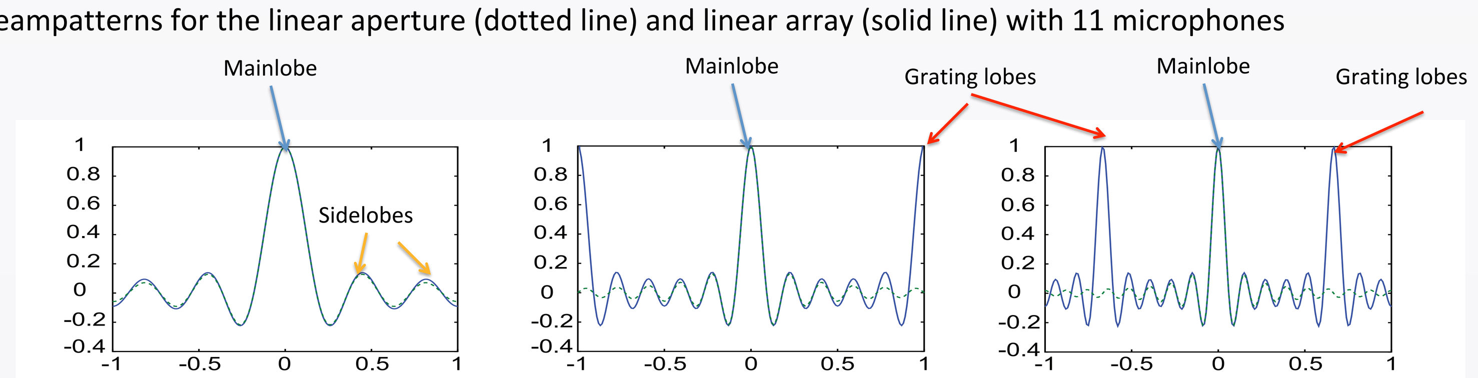


Linear Aperture



Beampattern plot (spatial directivity)

Beampatterns for the linear aperture (dotted line) and linear array (solid line) with 11 microphones



a) low frequency: $d/\lambda = fd/c = 1/2$ b) mid frequency: $d/\lambda = fd/c = 1$ c) high frequency: $d/\lambda = fd/c = 3/2$

- Unlike those for the linear aperture, beampatters for the linear array are periodic; there will be grating lobes because of spatial aliasing.
- We may not be able to pick up one direction at high frequency because of the grating lobes.
- The sidelobes limit the performance of interfering signal suppression.

Whatever method is used for estimating the weights of spatial filters, it will just control spatial directivity.

Speech enhancement approaches for spatial filter estimation

Processing Type	Need Adaptation Data?	Representative Methods
Real-time processing	No	Data-independent beamforming [1,2,3,4,7] Binaural processing [4,8] Beamforming with PIT [10]
	Yes	Adaptive optimum beamforming [1,3,4,7]
Batch processing	Yes	Maximum likelihood beamforming [6] Maximum super-Gaussian beamforming [3,4] Speech-noise mask-based beamforming [5] Source separation such as NMF [2,7] Black box approach such as deep clustering [5]

* Good speaker tracking will be required for real-time beamforming.

Most of conventional techniques are implemented in **BTK**: https://distantspeechrecognition.sourceforge.io/

References

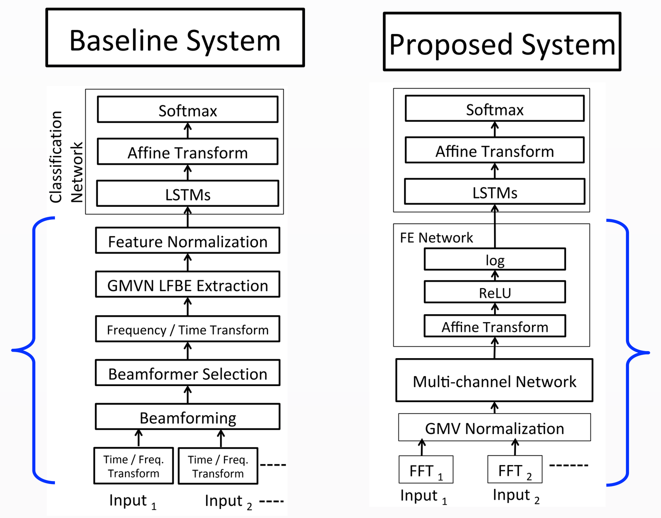
[1] T. M. Sullivan, *Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 1996.
[2] Omologo M., Matassoni M., Svaizer P. (2001) Speech Recognition with Microphone Arrays. in *Microphone Arrays*, Brandstein M., Ward D. (eds) Springer.
[3] M. Wölfel and J. W. McDonough, *Distant Speech Recognition*, Wiley, London, 2009.
[4] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*, John Wiley & Sons, 2012.
[5] S. Watanabe et al., "New Era for Robust Speech Recognition: Exploiting Deep Learning", 2018.
[6] M. L. Seltzer, B. Raj and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Trans. SAP* 2004.
[7] H. L. Van Trees, Optimum Array Processing, Wiley–Interscience, New York, 2002.
[8] R. M. Stern, DeL. Wang, and G. Brow, (2006). "Binaural Sound Localization," in *Computational Auditory Scene Analysis*, G. Brown and De. Wang (eds) Wiley/IEEE Press
[9] T. N. Sainath et al., "Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition", IEEE Trans. SLP, 2017
[10] T. Yoshioka et al., "Low-Latency Speaker-Independent Continuous Speech Separation", arxiv, 2019.

Our Far-field ASR system

- We unify a multi-channel front-end and phone classifier so as to minimize a phone classification error.
- Our initial fully-learnable network mimics a conventional ASR processing initially but removes a speech clean reconstruction step.

Hand-crafted front-end

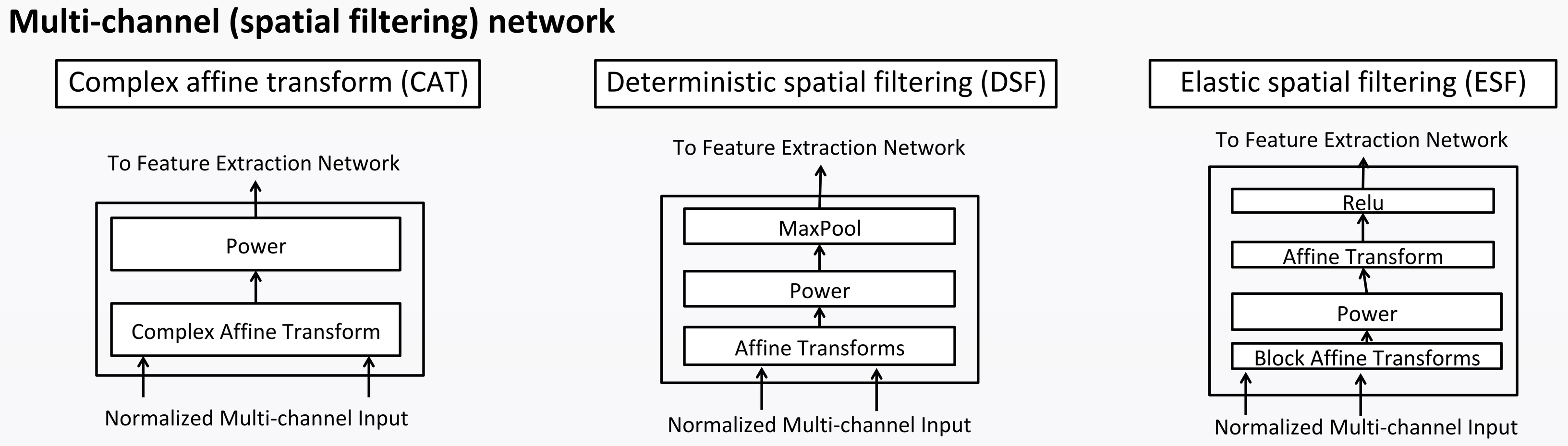
- Robust SD beamforming
- Beamformer selection
- LFBE feature extraction
- Causal feature normalization



Learnable Front-end optimized jointly

- 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 multi-channel (spatial filtering) layers initialized with super-directive (SD) beamformers' weights, we fine-tune the whole network with multi-channel DFT features.

Multi-channel (spatial filtering) network



Every multi-channel layer is initialized with SD beamformers' weights.

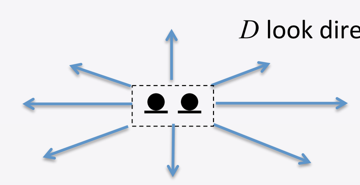
CAT is similar with the network architecture described in [9].

Relationship between beamforming and multi-channel network

Time-domain beamforming operation:

Assuming that we build D beamformers with S microphones, beamforming can be expressed as a convolution process of a multi-channel signal with D sets of FIR (or IIR):

$$\begin{bmatrix} y_1(t) \\ \vdots \\ y_D(t) \end{bmatrix} = \begin{bmatrix} \sum_{s=1}^{S} w_{1,s}(t) * x_s(t) \\ \vdots \\ \sum_{s=1}^{S} w_{D,s}(t) * x_s(t) \end{bmatrix}$$



This is normally implemented in (subband) frequency domain for the sake of computational efficiency.

Frequency-domain beamforming operation:

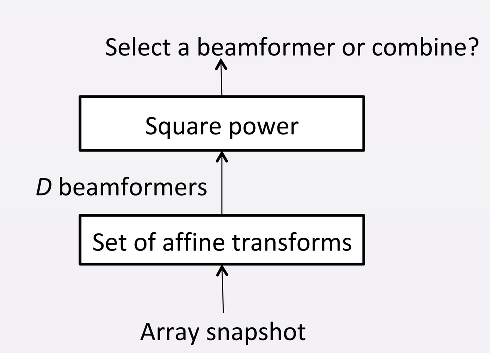
Frequency-domain beamforming at frequency f can be expressed with D x S complex linear transformation

$$\begin{bmatrix} Y_1(f,n) \\ \vdots \\ Y_D(f,n) \end{bmatrix} = \begin{bmatrix} W_{1,1}(f,n) & \cdots & W_{1,S}(f,n) \\ \vdots & \ddots & \vdots \\ W_{D,1}(f,n) & \cdots & W_{D,S}(f,n) \end{bmatrix} \begin{bmatrix} X_1(f,n) \\ \vdots \\ X_S(f,n) \end{bmatrix}$$

It is straightforward to build a neural network that is equivalent to multiple beamformers in the frequency domain.

D direction output for each frequency

$$\begin{bmatrix} Y_1(\omega_1) \\ \vdots \\ Y_D(\omega_1) \\ \vdots \\ Y_1(\omega_K) \\ \vdots \\ Y_D(\omega_K) \end{bmatrix} = \mathrm{pow}\left(\begin{bmatrix} \mathbf{w}^H(\omega_1, p_1)\mathbf{X}(\omega_1) \\ \vdots \\ \mathbf{w}^H(\omega_1, p_D)\mathbf{X}(\omega_1) \\ \vdots \\ \mathbf{w}^H(\omega_K, p_1)\mathbf{X}(\omega_K) \\ \vdots \\ \mathbf{w}^H(\omega_K, p_D)\mathbf{X}(\omega_K) \end{bmatrix} + \mathbf{b} \right)$$



Relationship between beamforming and source separation:

- Blind source separation (BSS) techniques attempt at unmixing multiple sound sources without any prior knowledge.
- For N active sources, BSS is formulated as:

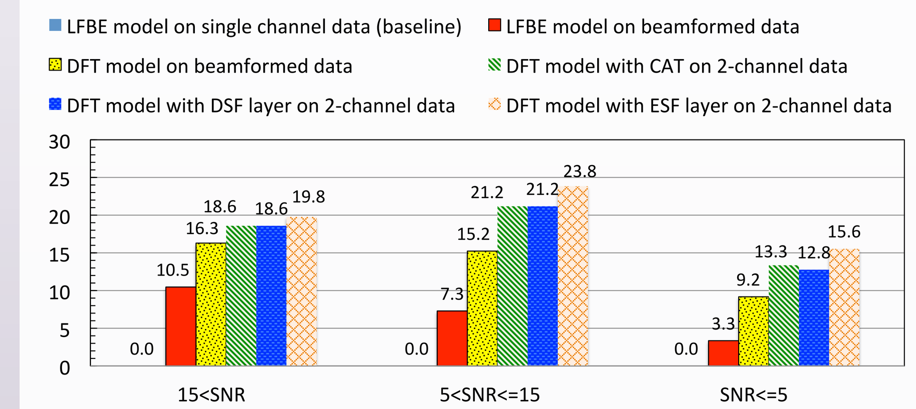
$$\begin{bmatrix} Y_1(f) \\ \vdots \\ Y_N(f) \end{bmatrix} = \begin{bmatrix} W_{11}(f) & \cdots & W_{1S}(f) \\ \vdots & \ddots & \vdots \\ W_{N1}(f) & \cdots & W_{NS}(f) \end{bmatrix} \begin{bmatrix} X_1(f) \\ \vdots \\ X_S(f) \end{bmatrix}$$

- BSS estimates the weights so as to minimize mutual information of each output; the LCMV adaptive beamformer can also do joint estimation with geometrical constraints; it is empirically known that the BSS solution for the row vector becomes null-steering beamformer's weights.

ASR Experiments

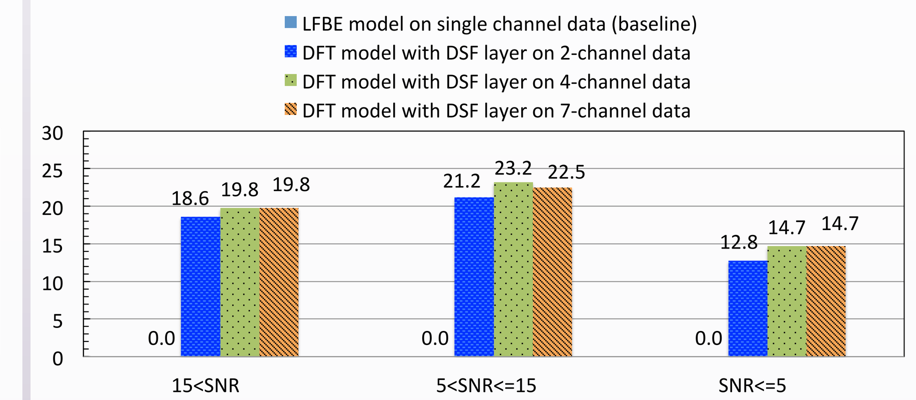
- We used approximately 1100 hours of speech spoken by human beings, collected with the 7 microphone circular array in various rooms and split 1,000 and 100 hours into training and test sets where there is no overlapping speaker between sets
- Part of data are captured through a Live traffic where the 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.

Overall Improvement from Baseline



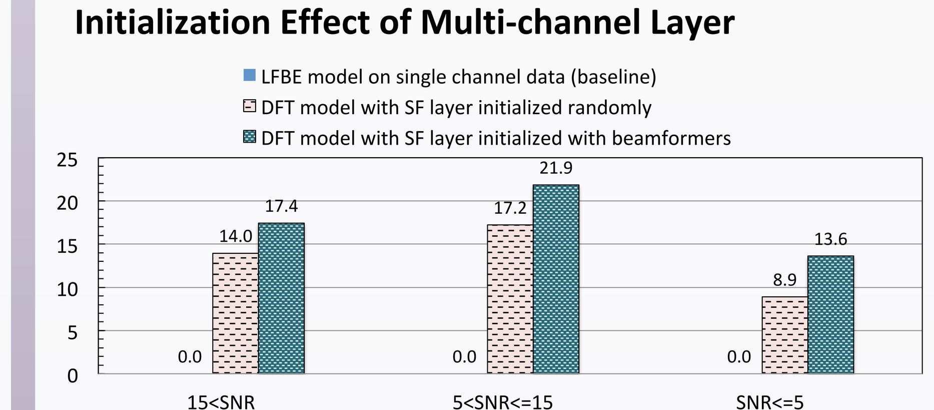
- Beamforming with 7 microphones can improve recognition performance.
- The fully-learnable two-channel models provide better accuracy than 7-channel beamforming.
- The ESF architecture provides the best accuracy in this experiment; the learnable feature front-end (DFT model) itself can improve recognition accuracy.

WER w.r.t a number of microphones



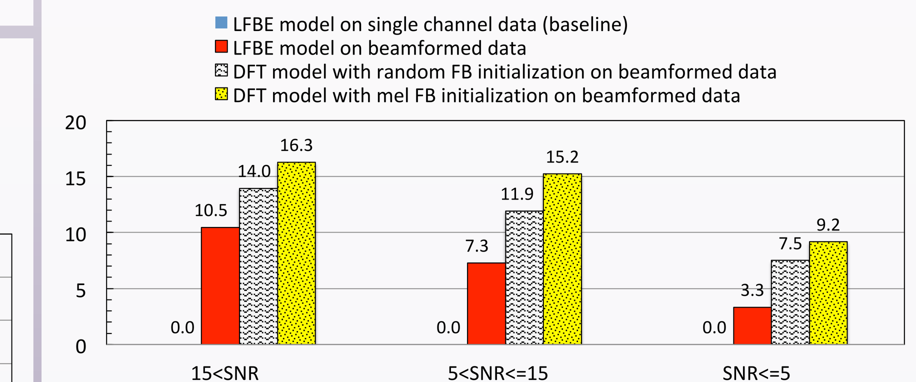
- Recognition accuracy is saturated at 4 microphones.
- There is a little degradation with 7 microphones, but this may change if more training data is used.

Initialization Effect of Multi-channel Layer



- Initializing the spatial filtering layer with beamformer's weight leads to better accuracy.

Effect of Learnable Feature Extraction Front-End



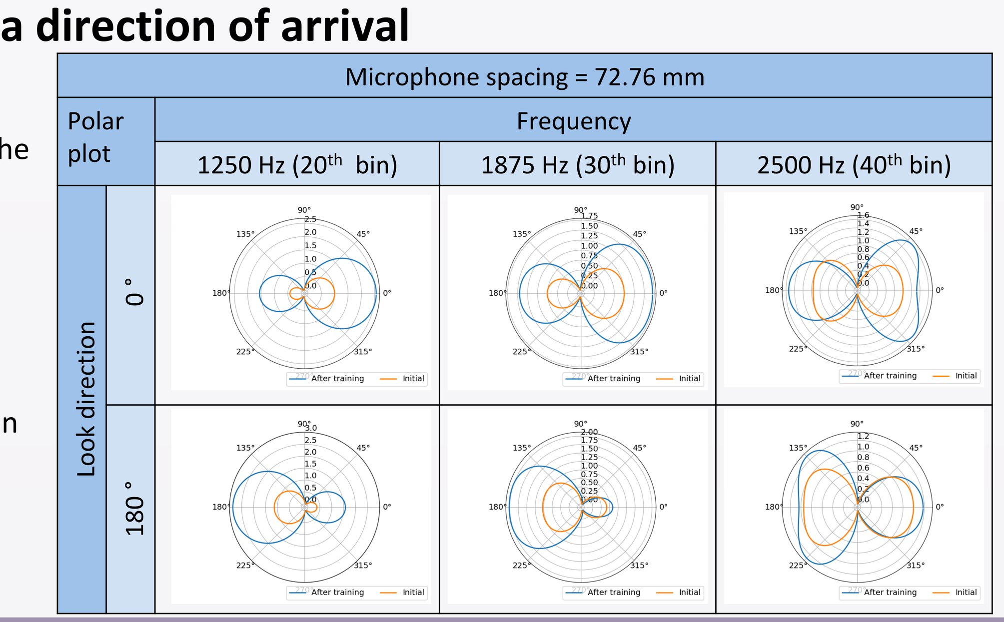
- The recognition accuracy can be improved by the learnable feature extraction network.
- The better accuracy is achieved by initializing the filter bank layer with mel-filter coefficients.

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

The left figure shows the SRP of super-directive beamforming (initial) and ESF network (after training) in the case of two-channel input.

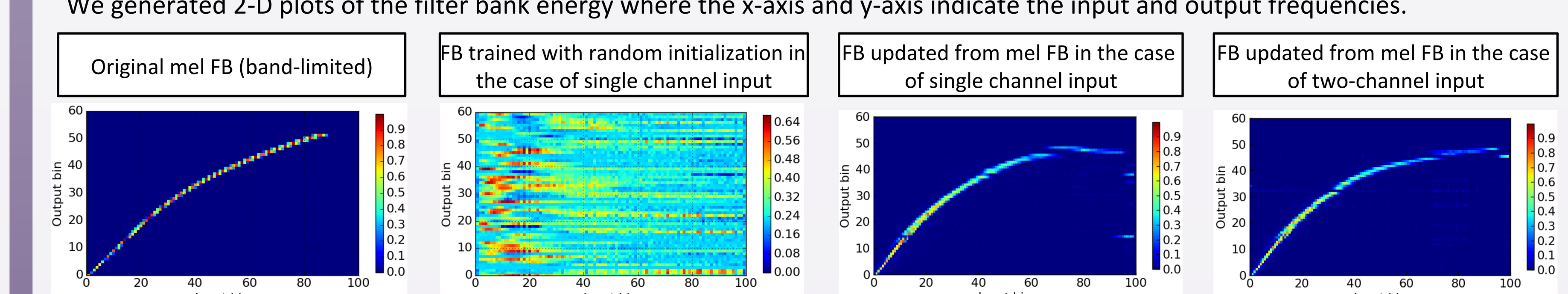
Each line indicates the directivity of the spatial filter, how much the filter strengthens or attenuates a signal coming from a particular direction.

The ESF network combines those filters with weights; Spatial filters (beamformers) were combined in a soft-decision manner so as to maximize the phone classification accuracy unlike determining a beam direction in a hard-decision manner.



Visualization of Learned Filter Bank (FB)

We generated 2-D plots of the filter bank energy where the x-axis and y-axis indicate the input and output frequencies.



- Random initialization provided a local minima solution.
- Initializing the affine transform with mel FB weights lead to a meaningful result, lower spectral resolution at a higher frequency.
- The number of input channels did not give an impact on filter bank estimate.

Conclusion

- The fully-learnable multi-channel AM can provide the better accuracy with a fewer sensors than classical beamforming.
- Everything can be learnt from a large amount of real data; we can avoid adaptation process that could hurt the performance.
- The learnable feature extraction front-end itself can provide better accuracy than the log mel-filter bank feature.
- Initializing the neural network with beamformer's weight and mel-filter coefficients leads to better recognition accuracy.