北京工業大學
BEIJINGUNIVERSITYOFTECHNOLOGY

A New Parametric Coding Method Combined Linear Microphone Array Topology

Presentation for Data compression conference

SASPL

Instructor： prof. Changchun Bao

Reporter： yao zhou

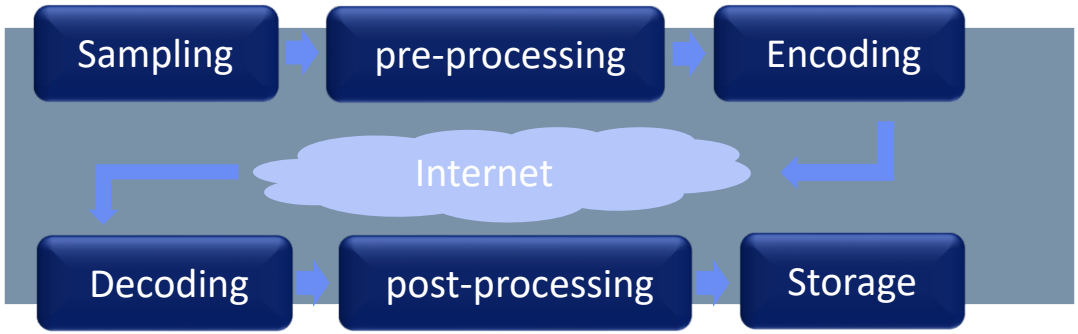
CONCENTS

Part 1

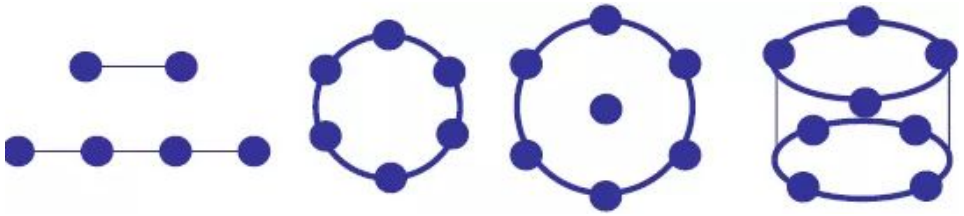
Backgrounds

This section introduced the application of spatial audio coding

Voice-over IP (VoIP) system



Sampling equipment:



Linear array

Planar array

Three-dimensional array

Multichannel speech coding scheme

Mono Codec
Mono Codec
Mono Codec

01010111...
01100111...
11010101...

Internet

Mono Codec
Mono Codec
Mono Codec

01010111...
01100111...
11010101...

Internet

Spatial Audio Coding (SAC) scheme

Downmix and Encoding

Parameter extraction and Quantization

01010111…

01100101…

Internet

Multi-channel Up-mix

Downmix signal decoding

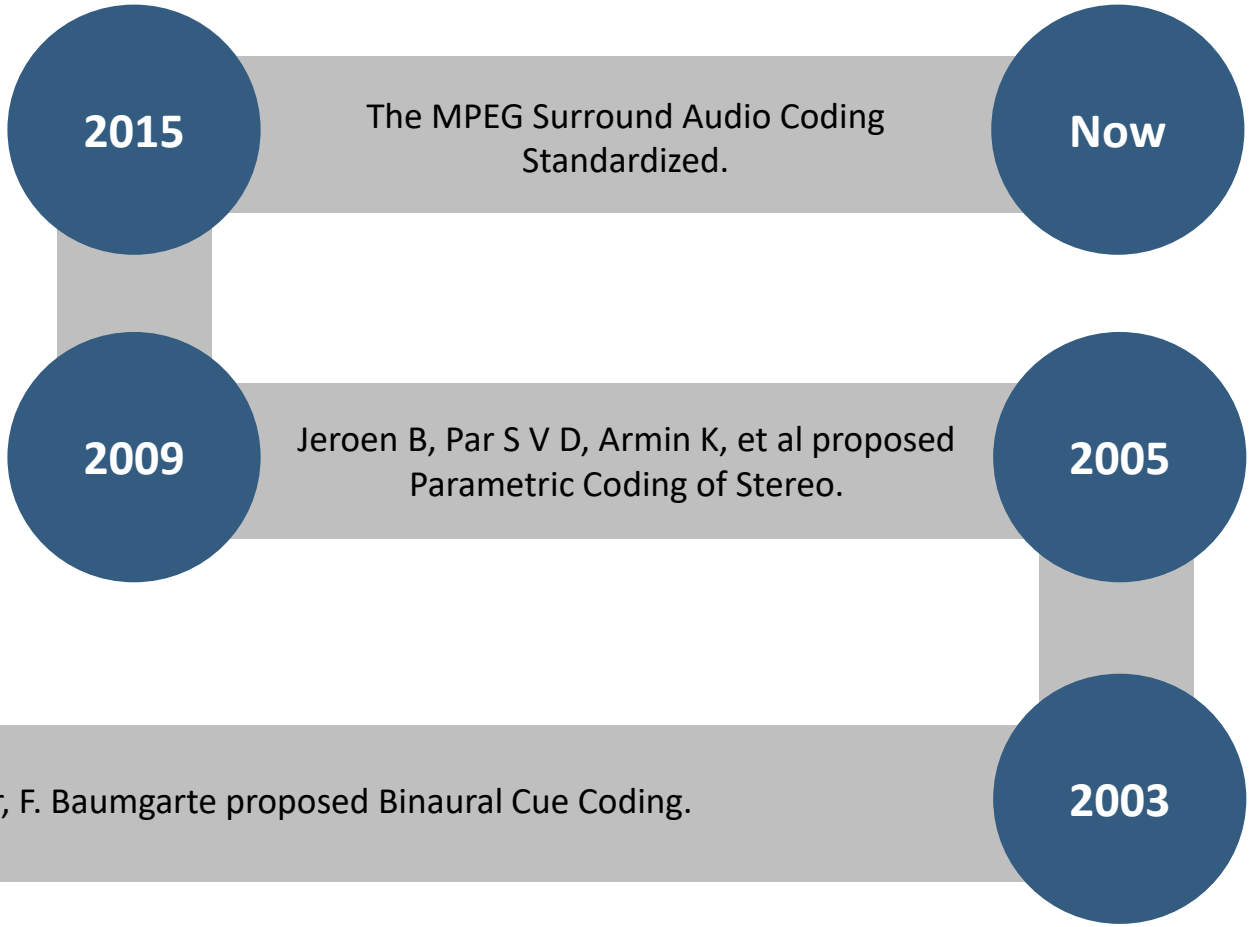
Parameter dequantization

01010111…

01100101…

Internet

SAC technique overview

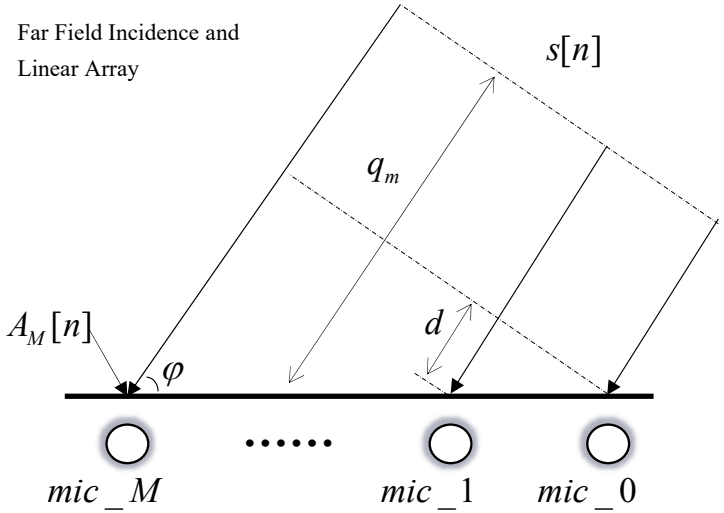


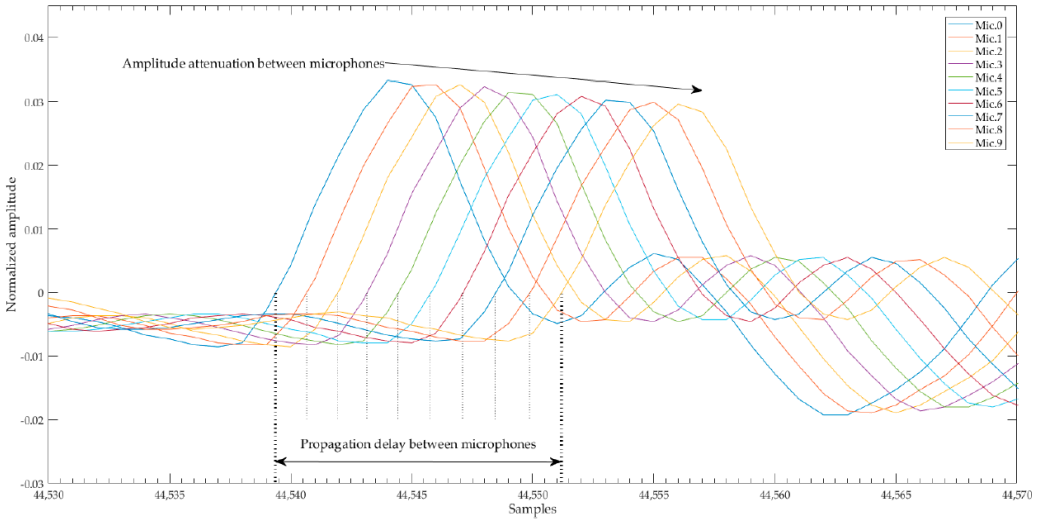
Part 2

Algorithm analysis

This part introduces the linear interpolation relationship between the signals of the linear microphone array

Far field linear microphone array





Signal model of far filed linear microphone :

$$x_m[n] = \frac{10}{\sqrt{4\pi q_m}} s\left[n - \frac{q_m}{v_c}\right] \qquad q_m = q_0 + m \times d$$

Linear energy decay gradient

Evenly distributed inter-channel time delay

Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. IEEE/ACM Trans. Audio Speech Lang. Process. **2017**, 25, 692–730.

Far field linear microphone array

$$x_m[n] = \frac{10}{\sqrt{4\pi q_m}} s\left[n - \frac{q_m}{v_c}\right] \qquad q_m = q_0 + m \times d$$

power of directional source:

$$\delta_s^2 = E[(\mathrm{s} - E[s])^2]$$
$$P_m = E[(x_m - E[x_m])^2]$$

Energy relationship between channels:

$$P_0 - P_m = \left(\frac{10}{\sqrt{4\pi q_0}}\right)^2 \delta_s^2 - \left(\frac{10}{\sqrt{4\pi(q_0 + md)}}\right)^2 \delta_s^2$$
$$= \left(\frac{10}{\sqrt{4\pi q_0}}\right)^2 \delta_s^2 \left(1 - \frac{1}{1 + \frac{2md}{q_0} + \frac{1}{4}\left(\frac{2md}{q_0}\right)^2}\right)$$

Considering Taylor Formula: $\frac{1}{1-\varepsilon} = 1 + \varepsilon + o(\varepsilon^2)$

$$P_0 - P_m = \left(\frac{10}{\sqrt{4\pi q_0}}\right)^2 \delta_s^2 \cdot \frac{2d}{q_0} m$$

$$\frac{P_0 - P_m}{m} = \frac{P_0 - P_M}{M}$$

The time delay relationship between channels are obvious:

$$f_m(\tau) = \frac{q_m - q_0}{v_c} = \frac{m \times d}{v_c} = \frac{m}{M}\left(f_M(\tau) - f_0(\tau)\right)$$

The coherence between channels are assumed to be the same in this situation.

Once the signals of first and last channels are obtained, the relationship between any intermedia channel and first channel could be derived.

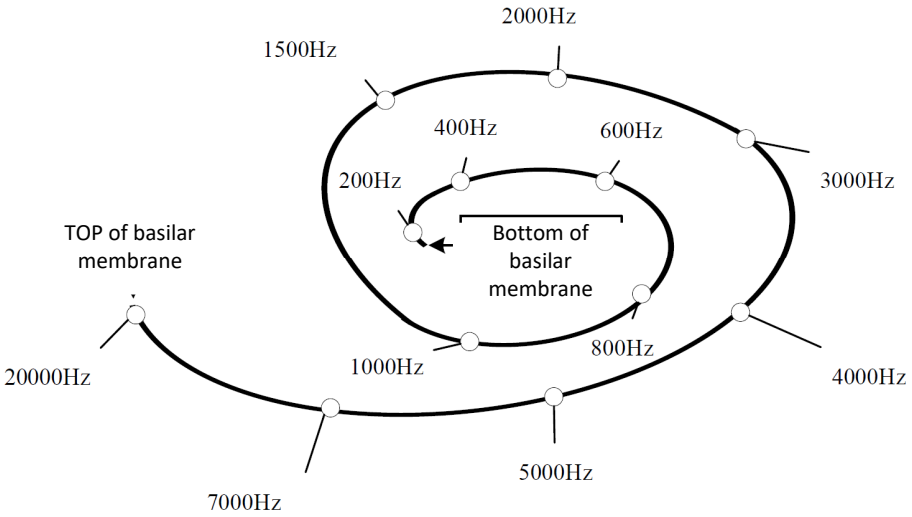
Part 3

Codec structure

This part introduces the encoder and decoder structure of the multichannel coding system

Human hearing system

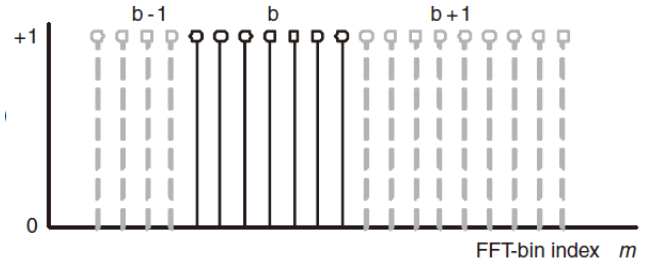
The human ear has different sensitivity to different frequency components of sound. (More sensitive to low frequency components than high frequency components)



$$\mathrm{ERB}(f_i) = 24.7 \times (4.37 \times f_i / 1000 + 1)$$

signals → Window → FFT → Group

1. Signals are segmented into overlapping frames with a window ;
2. Each segment is subsequently transformed to the frequency domain using an FFT;
3. Frequency domain signals are divided into nonoverlapping subbands by grouping of FFT bins. The form is following equivalent rectangular bandwidth (ERB).



Multichannel encoder

1. the downmix signal is obtained by averaging all signals time-aligned with the first channel, then sent to any Mono-channel audio codec for encoding and transforming.

2. Extract the inter-channel level difference(ICLD) parameters:

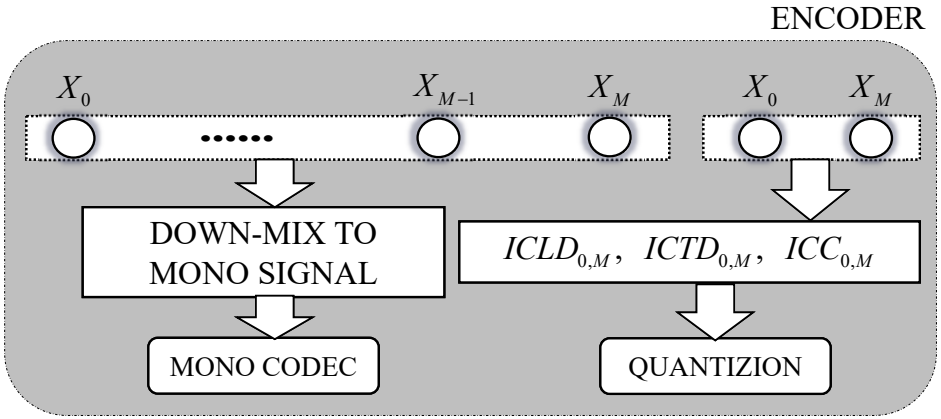
$$ICLD_{0,M}[i,b] = 10\log_{10}\frac{\sum_{k}^{N-1}|X_M[i,k]|^2}{\sum_{k}^{N-1}|X_0[i,k]|^2}$$

3. Extract the inter-channel time difference(ICTD) parameters:

$$ICTD_{0,M}[i,b] = \underset{\tau}{\operatorname{argmin}}\left(\sum_{k}^{N-1}(\angle(X_0^*[i,k]X_M[i,k]) - \frac{2\pi k}{N}\tau)\right)$$

4. Extract the inter-channel coherence difference(ICC) parameters:

$$ICC_{0,M}[i,b] = \frac{R\left(\sum_{k}^{N-1}(X_0^*[i,k]X_M[i,k])\right)}{\sqrt{\sum_{k}^{N-1}|X_0[i,k]|^2\sum_{k}^{N-1}|X_M[i,k]|^2}}$$



ICLD: the energy level ratio between channels

ICTD: the time difference between channels

ICC: the correlation or coherence between channels.

Multichannel decoder

1 Use the received downmix signal after transmission and a unique set of spatial parameters to up-mix to obtain the reconstructed first and last channel signals:

$$\begin{bmatrix} \tilde{x}_0 \\ \tilde{x}_M \end{bmatrix} = \begin{bmatrix} \lambda_1 e^{i\theta_1}\cos(\alpha) & \lambda_1 e^{i\theta_1}\sin(\alpha) \\ \lambda_2 e^{i\theta_2}\cos(-\alpha) & \lambda_2 e^{i\theta_2}\sin(-\alpha) \end{bmatrix} \begin{bmatrix} \tilde{c} \\ D(\tilde{c}) \end{bmatrix}$$

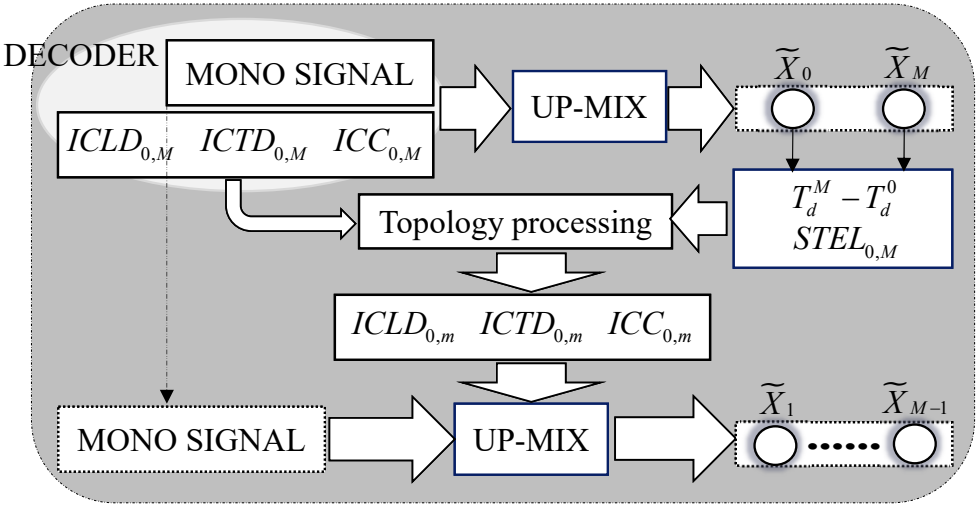
$\lambda_1, \lambda_2, \theta_1, \theta_2, \alpha$ are correlated with transmitted spatial parameters and $D(\tilde{c})$ is a decorrelation process for downmix signal, all above is detailed in reference below.

2 Interpolating other the spatial parameters of intermedia channels and the first channel:

$$ICLD_{0,m}[i, b] = 10\log_{10}\frac{\tilde{P}_0[i]}{\tilde{P}_M[i]} - \frac{m}{M}\left(\frac{\tilde{P}_0[i]}{\tilde{P}_M[i]} - 1\right) + ICLD_{0,M}[i, b]$$

$$ICTD_{0,m}[i, b] = \frac{m}{M}\left(ICTD_{0,M}[i, b]\right)$$

$$ICC_{0,m}[i, b] = ICC_{0,M}[i, b]$$



3 When the decoded downmix signal and spatial parameters between the channels are available.
the signals of each channel are up-mixed as same as the first step.

Breebaart, J., van de Par, S., Kohlrausch, A. *et al.* Parametric Coding of Stereo Audio. *EURASIP J. Adv. Signal Process.* 2005, 561917 (2005).

Part 4

Evaluations

This part illustrate the experiment settings and metrics evaluations.

Experiments

Basic Settings

Based on MASS method and Pyacousticroom simulator, a shoebox room is established, and then simulated data is generated. (Reference below)

TIMIT → **linear array**

8 microphone array, spaced 4 cm apart

Cheng, Rui & Bao, Changchun & Cui, Zihao. (2020). MASS: Microphone Array Speech Simulator in Room Acoustic Environment for Multi-Channel Speech Coding and Enhancement. Applied Sciences. 10. 1484.

Room Simulation — Pyroomacoustics 0.6.0 documentation

https://github.com/LCAV/pyroomacoustics

&

Bitrate settings

	Down-mix	Parameters	Total
PS	512 kb/s	40 kb/s	552 kb/s
BCC	128 kb/s	70 kb/s	198 kb/s
Proposed	128 kb/s	10 kb/s	138 kb/s

Explanation：

Downmix signals are transmitted by EVS with bitrate 128Kbps.

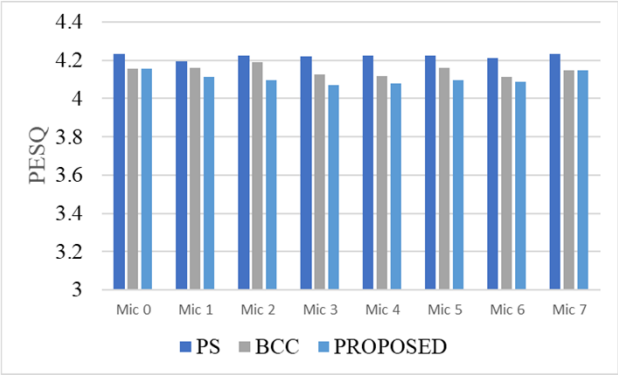
1. PS denotes Parametric stereo method
   It produces 4 sets of downmix signals and parameters
2. BCC denotes Binaural cue coding method
   It produces 1 set of downmix signals and 7 sets of parameters
3. Proposed denotes topology combined method
   It produces 1 set of downmix signals and1 set of parameters

Evaluations

Perceptual evaluation of speech quality (**PESQ**)

A kind of objective speech quality evaluation method.

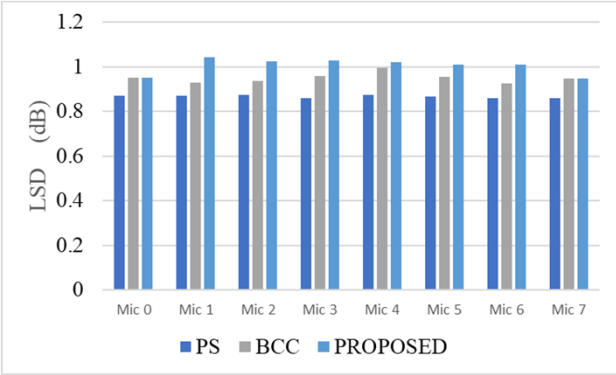
The score ranges from -0.5 to 4.5, the higher, the better.



log spectral distance(**LSD**)

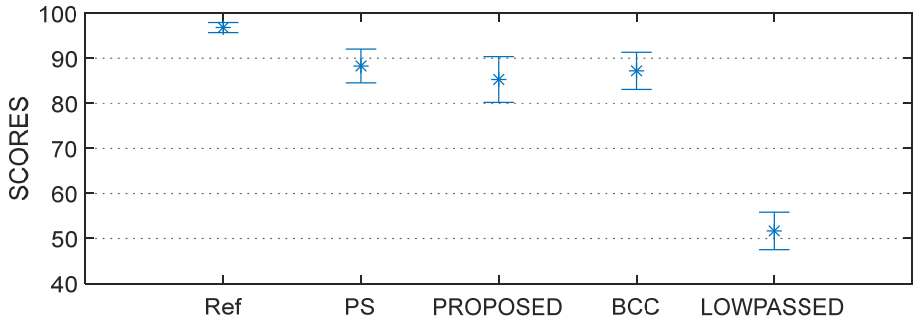
A kind of objective speech quality evaluation method.

It measures the spectral distortions, the lower, the better.



ITU-R Recommendation BS.1534 (MUSHRA)

It is a kind of subjective tests, which is designed to compare the audio quality of several test conditions with intermediate impairments to a high-quality reference. Scores are the higher, the better.



Part 5

Conclusion

This part makes a conclusion about this work

Conclusion

SAC

This paper follows the concept of spatial audio coding (SAC), constructed a multichannel coding system for a linear microphone array. **The coding efficiency is improved by transmitting the spatial parameters rather than all channels.**

Topology

The method proposed only needs to transmit one down-mix signal and the **minimum spatial parameters** related to the first and last channels of the array. **By combining with the topology of the array, finally all the array signals are recovered.**

Further

It is potential to consider the feasibility of using array topology to further reduce the redundancy and improve the effectiveness of multi-channel signal transmission **in a variety of microphone array structures**, not only the linear array.

北京工業大學

BEIJINGUNIVERSITYOFTECHNOLOGY

Thanks

That is all my presentation

SASPL

Instructor： prof. Changchun Bao

Reporter： yao zhou