Binaural tracking of multiple moving sources

Speech Signal Processing EE627

Professor :- Rajesh M. Hegde

TA:- Vishnuvardhan V

Participants:-

1. Rishabh Yadav (14554)
2. Ravish Raj (14542)
3. Vartak Sahil Dileep (14789)

Source: Technical Report OSU-CISRU-4/06-TR44

Objective

- Tracking of multiple moving sources using binaural input.

Methodology

- Algorithm integrates probabilities across reliable frequency channels in order to produce likelihood function in target space, which describes the azimuths of active sources at particular time frames
- Hidden Markov Model(HMM) is employed to form continuous tracks and automatically detect the number of active sources across time
- Addressing auditory scene analysis with moving sound sources.

Definitions

1. **ITD**: The **interaural time difference** when concerning humans or animals, is the difference in arrival time of a sound between two ears. It is important in the localization of sounds, as it provides a cue to the direction or angle of the sound source from the head.
2. **IID**: **Interaural Intensity Difference** (IID) or Interaural Level Difference (ILD); Sound from the right side has a higher level at the right ear than at the left ear, because the head shadows the left ear.
3. **HRTF**: A **head-related transfer function (HRTF)** also sometimes known as the *anatomical transfer function* (ATF) is a response that characterizes how an ear receives a sound from a point in space

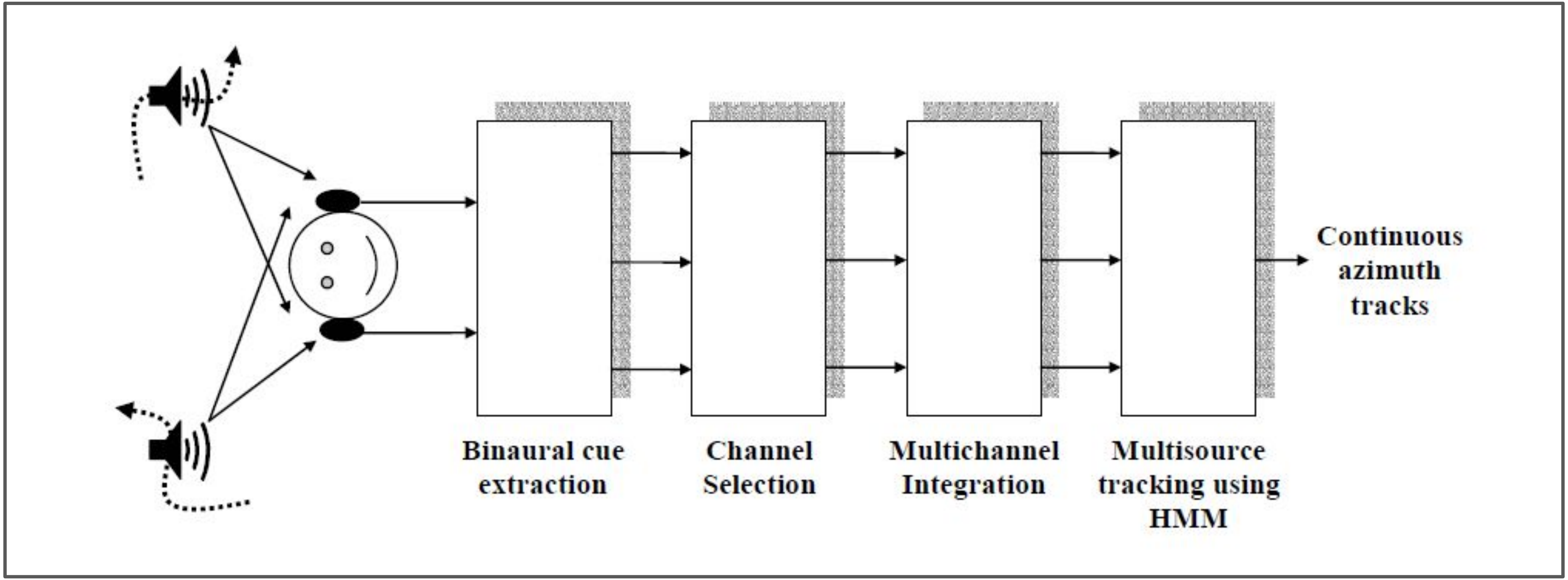
Model Architecture Stages

1) a model of the auditory periphery and binaural cue estimation

2) a channel selection mechanism that identifies reliable frequency channels in each time frame;

3) a multichannel statistical integration method that produces the likelihood function for target subspaces

4) a continuous HMM model for multi-source tracking.



A schematic diagram of the proposed multi-source tracking system

1. The input to their model is a binaural response of a KEMAR(Knowles Electronics Manikin for Acoustic Research) dummy head to an acoustic scene with multiple moving sources.
2. For each frequency channel, normalized cross-correlation functions between the two ear signals are computed in consecutive time frames. The time lag of a peak in the cross-correlation function is a candidate for **ITD** estimation. Ambiguity at high frequencies can be resolved by using IID information.
3. Channel reliability is measured by height of peak in cross-relation function since channel selection is our second stage.
4. Considering the statistical distribution of the ITD-IID estimates, We formulate the probability of each channel supporting the hypothesis and then employ an integration method to produce the likelihood of observing the configuration.
5. Proposing an HMM model that allows jumping between subspaces within each of which only a subset of the total number of sources is active which combines the likelihood model for the dynamics of source motion and jump probabilities which leads to optimal azimuth tracks using the Viterbi decoding algorithm.

Modeling Auditory Motion

1. Changes in ITD and IID would provide velocity info and perceive and track the changing source location.
2. The HRTF Catalog provides 256 point impulse responses for a fixed number of locations residing on a 1.4 m radius sphere around the KEMAR head.
3. The resolution in the horizontal plane is 5^o azimuth. The sampling rate is fixed at 44.1 kHz.
4. HRTF is divided into a cascade of a minimum-phase filter and a pure delay line.
5. The motivation is that minimum-phase systems behave better than the raw measurements for interpolation both in the phase and the magnitude response.
6. The impulse response of an arbitrary direction of sound incidence is obtained by interpolating separately the minimum-phase filters and the time delays.

Binaural Processing

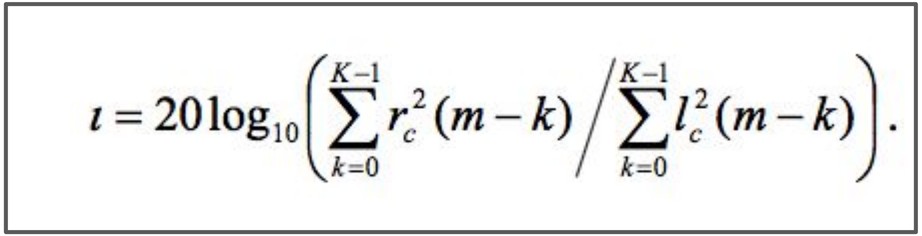
- ITD information is extracted by employing normalized cross-correlation function computed at different lags.

$$C(c,m,\tau)=\frac{\sum_{k=0}^{K-1}(l_c(m-k)-\overline{l}_c)(r_c(m-k-\tau)-\overline{r}_c)}{\sqrt{\sum_{k=0}^{K-1}(l_c(m-k)-\overline{l}_c)^2}\sqrt{\sum_{k=0}^{K-1}(r_c(m-k-\tau)-\overline{r}_c)^2}},$$

For frequency channel c, time frame m, and lag て.

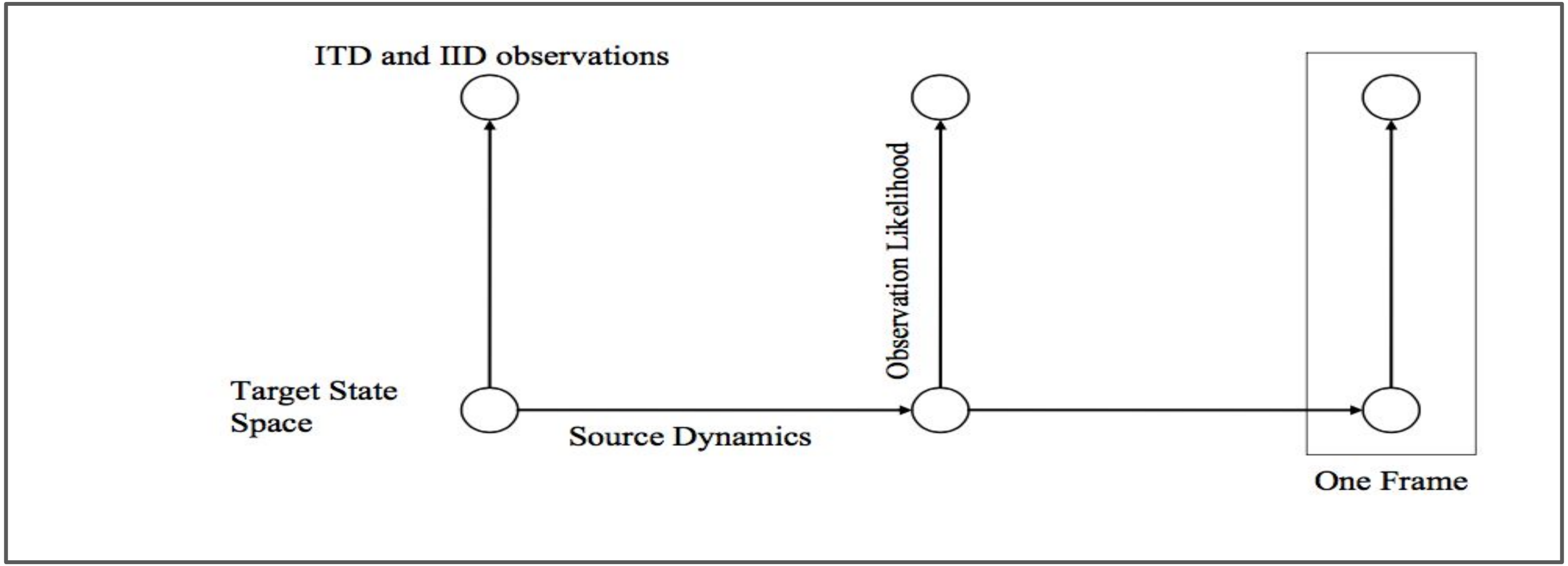
l_c,r_c are the left and right peripheral output for channel c.

- IID information is extracted for frequency channel c and time frame m by computing the energy ratio at the two ears, expressed in decibels:



$$\iota = 20\log_{10}\left(\sum_{k=0}^{K-1} r_c^2(m-k) \Big/ \sum_{k=0}^{K-1} l_c^2(m-k)\right).$$

Proposed HMM Framework



- HMM is completely defined by the following: 1) the possible target state space; 2) the transition probabilities that reflect the evolution of the target states across time; and 3) the observation probabilities conditioned on the target states, also known as the observation likelihood.

Points about HMM framework

1. A state in the target space specifies what the active sources are as well as their azimuth information at a particular time frame.
2. The target space is decomposed into subspaces; each subspace corresponds to a subset of active sources.
3. The transition probability between states in neighboring time frames must take into account both the jump probability between subspaces and the temporal evolution within individual subspaces.
4. Finally, a statistical model that integrates ITD and IID observations in different frequency channels is used to construct the observation likelihood in the target space.
5. To increase the validity of the system only frequency channels that are dominated by **a single source** and thus deemed reliable are considered in our statistical integration.

Dynamics Model

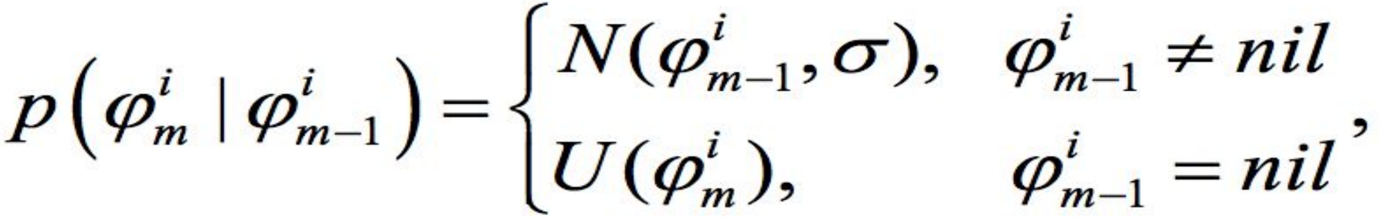
$$S = S_0 \cup S_1^1 \cup S_1^2 \cup S_1^3 \cup S_2^{1,2} \cup S_2^{1,3} \cup S_2^{2,3} \cup S_3 ,$$

- S is defined as our target state space and S_0 is the silence space with no active source, S^i_1 is the state space for a single active source i, $S_2^{i,j}$ is the state space for two simultaneously active sources i and j, and S_3 is the state space for all three active sources.

$$p\left(\boldsymbol{x}_m, s_m \mid \boldsymbol{x}_{m-1}, s_{m-1}\right) = p(s_m \mid s_{m-1}) \prod_{i \in I} p\left(\varphi_m^i \mid \varphi_{m-1}^i\right),$$

- Suppose that the state of the system at frame m, $x_m=(\Phi^1_m,\Phi^2_m,\Phi^3_m)$ is in the subspace s_m and the sources are independent of each other. Then the state transitions are described as above where where $p(s_m \mid s_{m-1})$ is the jump probability between subspaces, I is the set of active sources at time frame m, and $p(\Phi^i_m \mid \Phi^i_{m-1})$gives the temporal evolution of the ith source.

Assuming that an active source moves slowly and follows a linear trajectory with additive Gaussian noise. Also, when a source transitions from silence to activity we assume a uniform distribution in the azimuth space. Therefore the dynamics of the ith source is described by :



$$p\left(\varphi_m^i \mid \varphi_{m-1}^i\right) = \begin{cases} N(\varphi_{m-1}^i, \sigma), & \varphi_{m-1}^i \neq nil \\ U(\varphi_m^i), & \varphi_{m-1}^i = nil \end{cases},$$

where nil stands for silence, N(ϕ,σ) denotes the Gaussian distribution with mean ϕ and standard deviation σ which is set to a small value. U denotes the uniform distribution in the azimuth range [-90°, 90°].

Statistics of ITD and IID

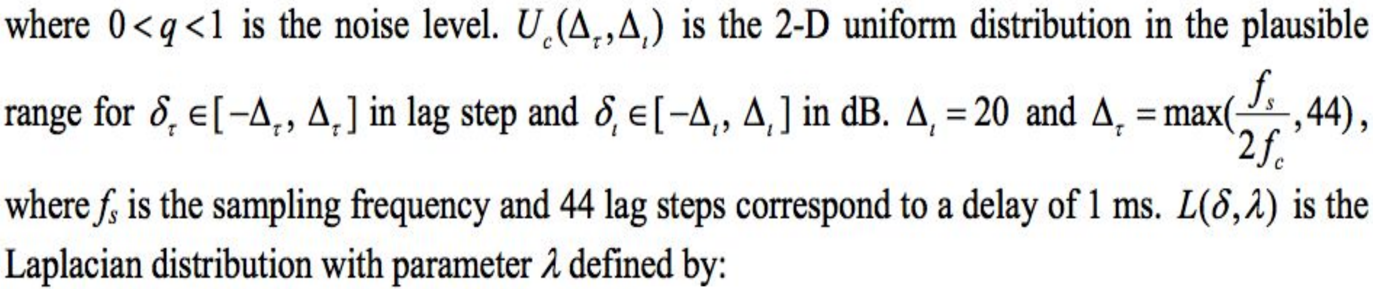
- For a single source the normalized cross-correlation has a maxima of 1 when the left and right signal are identical except for a time delay and an intensity difference.
- But for more than one source at different location only those peaks which cross the threshold are selected.
- The estimated ITD and IID signal a specific source location.
- For each frequency channel, the reference ITD are obtained from simulated white noise signals at different azimuth angles.
- Then ITD and IID deviations are calculated and we define the joint distribution of ITD and IID deviations in channel c.

$$\boxed{\begin{aligned}\delta_\tau &= \tau - \tau_{ref}(c,\varphi),\\ \delta_\iota &= \iota - \iota_{ref}(c,\varphi),\end{aligned}}$$

← **Deviations**

$$p_c(\delta_\tau,\delta_\iota) = (1-q)L(\delta_\tau,\lambda_\tau(c))L(\delta_\iota,\lambda_\iota(c)) + qU_c(\Delta_\tau,\Delta_\iota),$$

← **Joint Distribution**



where $0<q<1$ is the noise level. $U_c(\Delta_\tau,\Delta_\iota)$ is the 2-D uniform distribution in the plausible range for $\delta_\tau \in [-\Delta_\tau,\, \Delta_\tau]$ in lag step and $\delta_\iota \in [-\Delta_\iota,\, \Delta_\iota]$ in dB. $\Delta_\iota = 20$ and $\Delta_\tau = \max(\frac{f_s}{2f_c}, 44)$, where f_s is the sampling frequency and 44 lag steps correspond to a delay of 1 ms. $L(\delta,\lambda)$ is the Laplacian distribution with parameter λ defined by:

Laplaceian Distribution →

$$L(\delta,\lambda) = \frac{1}{2\lambda}\exp\left(-\frac{|\delta|}{\lambda}\right).$$

Likelihood Model

- We derive the conditional probability density $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$, often referred to as the likelihood, which statistically describes what a single frame of ITD and IID observations relate to the joint state $\boldsymbol{x}$ of the source locations to be tracked. T_c is the set of time lags $\boldsymbol{\tau}_c$ corresponding to the local peaks in the cross-correlation function and ιc is the estimated IID for channel c.
- the conditional probability $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$, for the one-source subspaces, i.e. $\boldsymbol{x} \in S_1^1 \cup S_1^2 \cup S_1^3$

- the conditional probability of the observations in channel c with respect to the one-source state x is given by:

$$p(T_c, \iota_c \mid \boldsymbol{x}) = \begin{cases} p_c(\delta_\tau, \delta_\iota), & \textit{if channel c is selected} \\ qU_c(\Delta_\tau, \Delta_\iota), & \textit{else} \end{cases},$$

- The observation probability in the current time frame conditioned on the one-source state x is smoothed using a root operation:-

$$p(\{T_c, \iota_c\} \mid \boldsymbol{x}) = \kappa \sqrt[N_b]{\prod_c p(T_c, \iota_c \mid \boldsymbol{x})},$$

where N_b=20 is the root number and κ is a normalization factor.

Conditional Probability for Kth Source

$$p(T_c, \iota_c \mid \boldsymbol{x}, k) = \begin{cases} qU_c(\Delta_\tau, \Delta_\iota), & \textit{if channel c not selected} \\ p_c(\delta_\tau^k, \delta_\iota^k), & \textit{if channel c belongs to source k}, \\ \max[p_c(\delta_\tau^1, \delta_\iota^1), p_c(\delta_\tau^2, \delta_\iota^2)], & \textit{else} \end{cases}$$

- Finally, the conditional probab $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$, for the current time frame is the larger of assuming either the first or the second hypothesized source to be the stronger source:

$$p(\{T_c, \iota_c\} \mid \boldsymbol{x}) = \alpha_2 \max[p(\{T_c, \iota_c\} \mid \boldsymbol{x}, 1), p(\{T_c, \iota_c\} \mid \boldsymbol{x}, 2)],$$

THANK YOU!