Columbia University

LabROSA
Laboratory for the Recognition and Organization of Speech and Audio

Deformable Spectrograms

Manuel Reyes , Dan Ellis & Nebojsa Jojic ; Columbia University , Microsoft Research

Microsoft Research

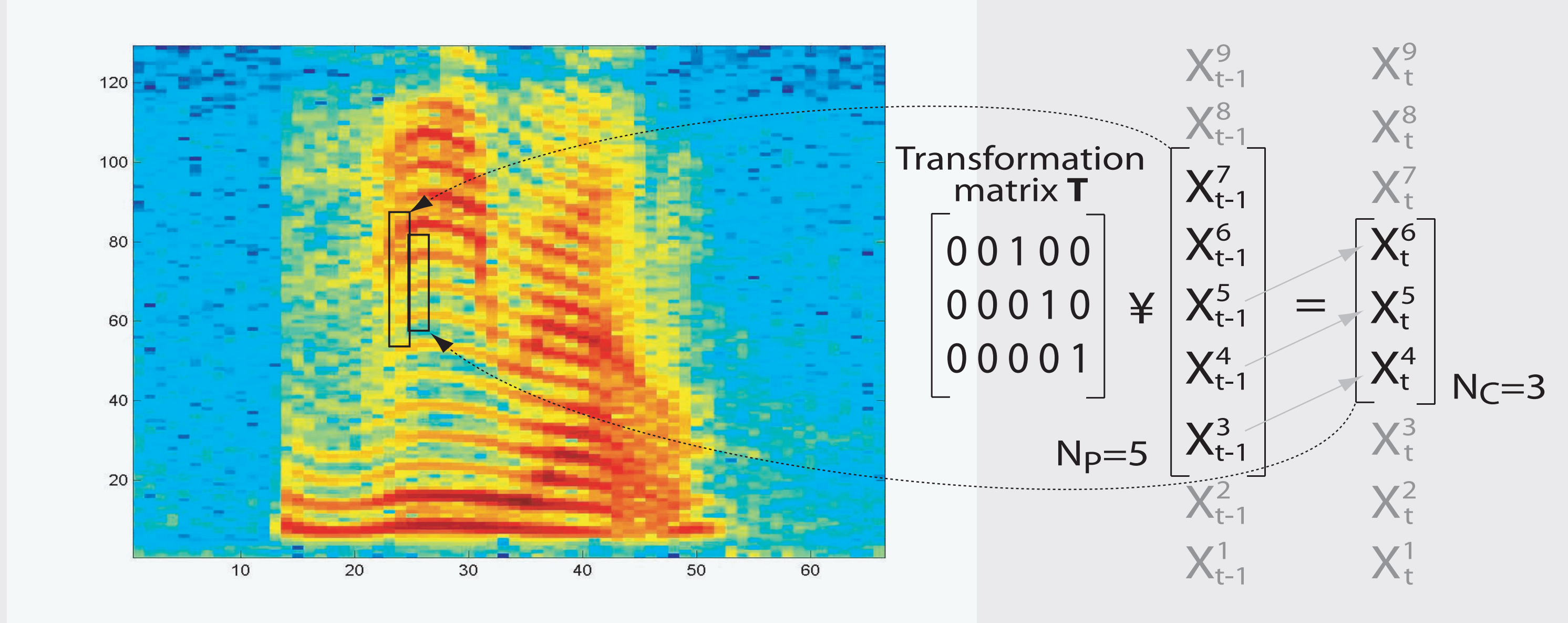
Summary: Our model focuses on local deformations of adjacent bins in a time-frequency surface to explain an observed sound, using explicit representation only for those bins that cannot be predicted from their context.

Introduction

We propose a model that focuses on local deformations of adjacent bins in a time-frequency surface to explain an observed sound, using explicit representation only for those bins that cannot be predicted from their context. The idea is to capture the self-similarity and dynamics of an unoccludded speech signal, such that those characteristics could later be exploited to separate occluded regions, when overlaps with other sources are encountered.

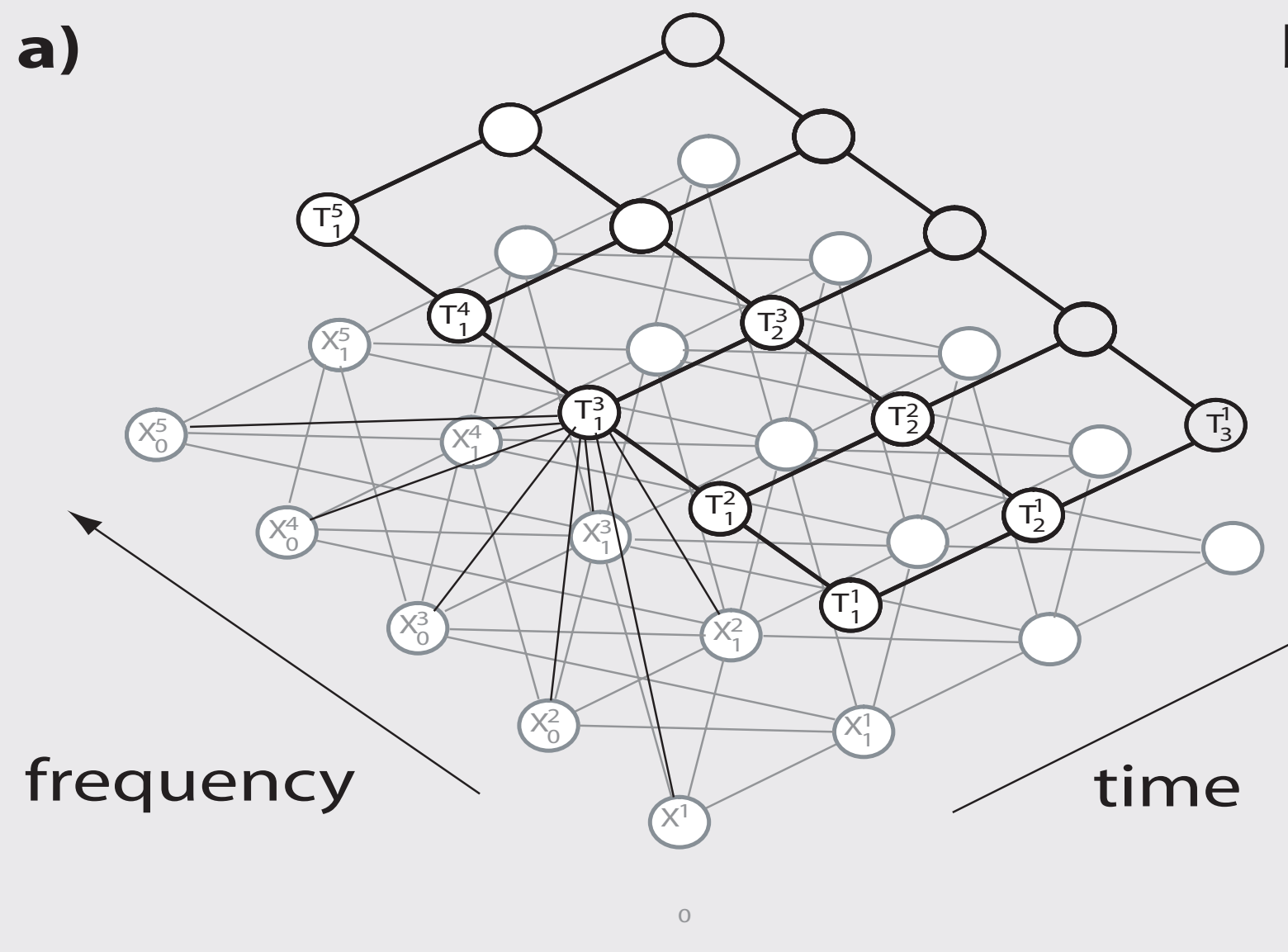
The transformation model.

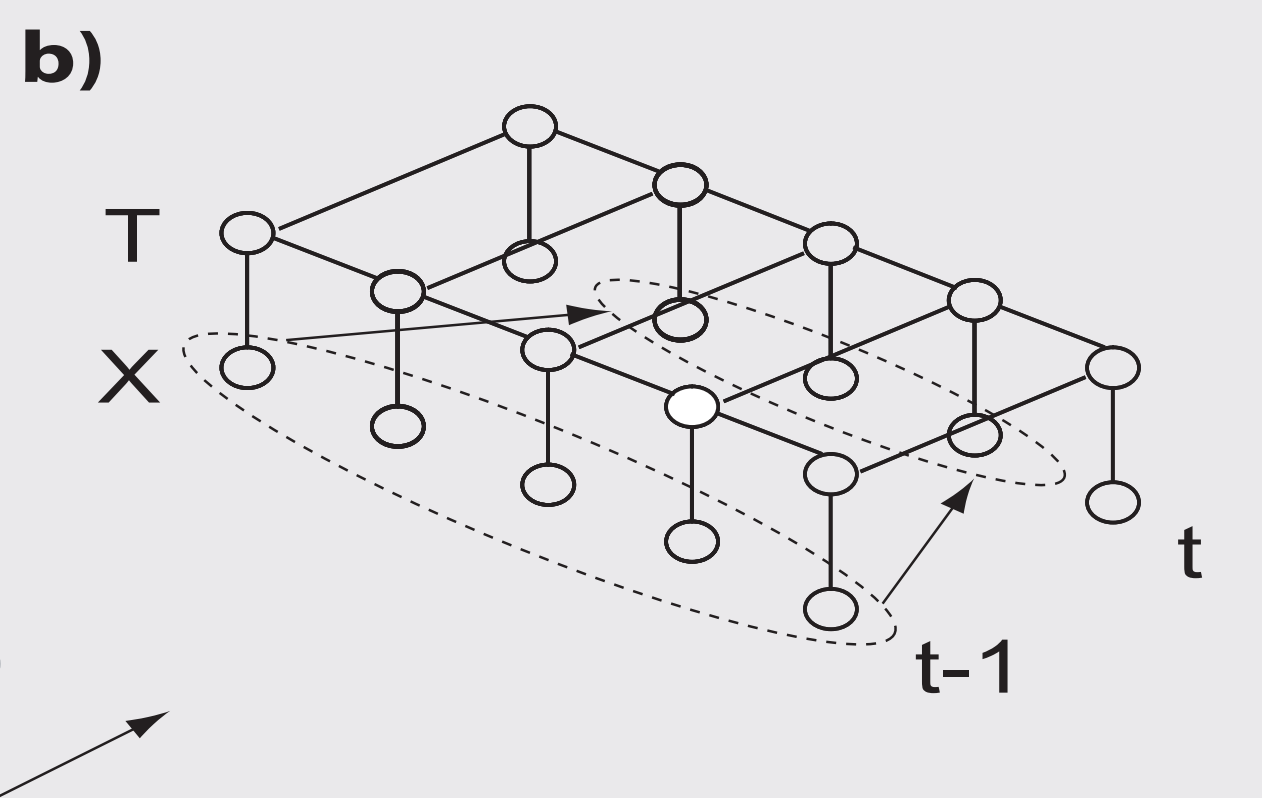
A patch of N_1 frequency bins, center at the k_{th} band from frame t is generated from a "transformation" of a N_2 frequency bins patch center at the k_{th} from frame t-1.


120
100
80
60
40
20
10
20
30
40
50
60
Transformation matrix **T**
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1
¥
=
X^9_{t-1} X^8_{t-1} X^7_{t-1} X^6_{t-1} X^5_{t-1} X^4_{t-1} X^3_{t-1} X^2_{t-1} X^1_{t-1}
X^9_{t} X^8_{t} X^7_{t} X^6_{t} X^5_{t} X^4_{t} X^3_{t} X^2_{t} X^1_{t}
N_P=5
N_C=3

$$p(X_t^{[k-n1,k+n1]} | X_t^{[k-n2,k+n2]}, T_t^k) = N(X_t^{[k-n1,k+n1]}; T_t^k X^{[k-n2,k+n2]}, \Phi^{[k-n1,k+n1]})$$

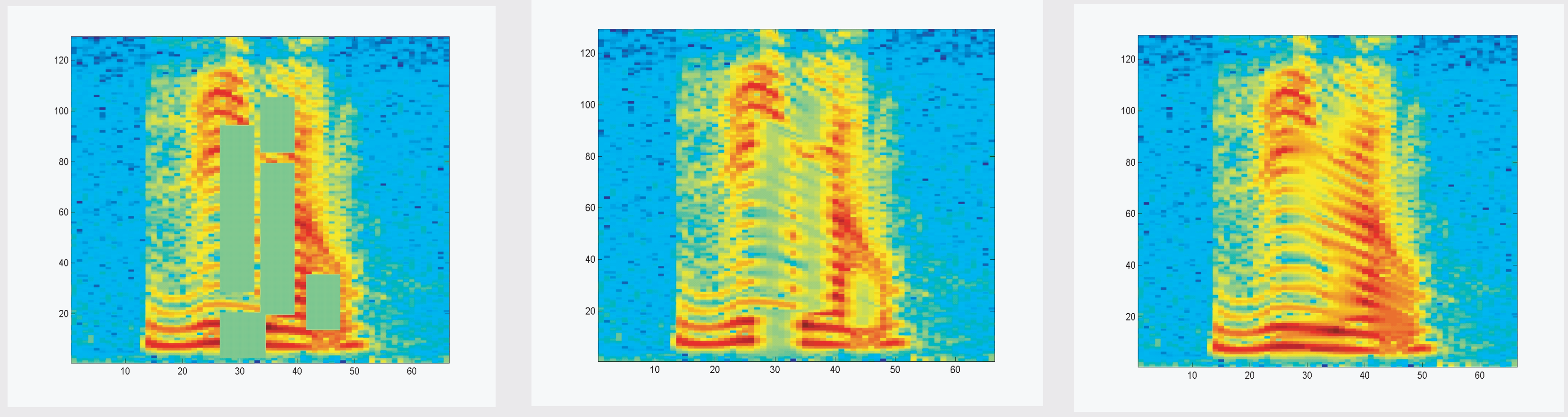
The Graphical Model.





Inference done using loopy belief propagation.

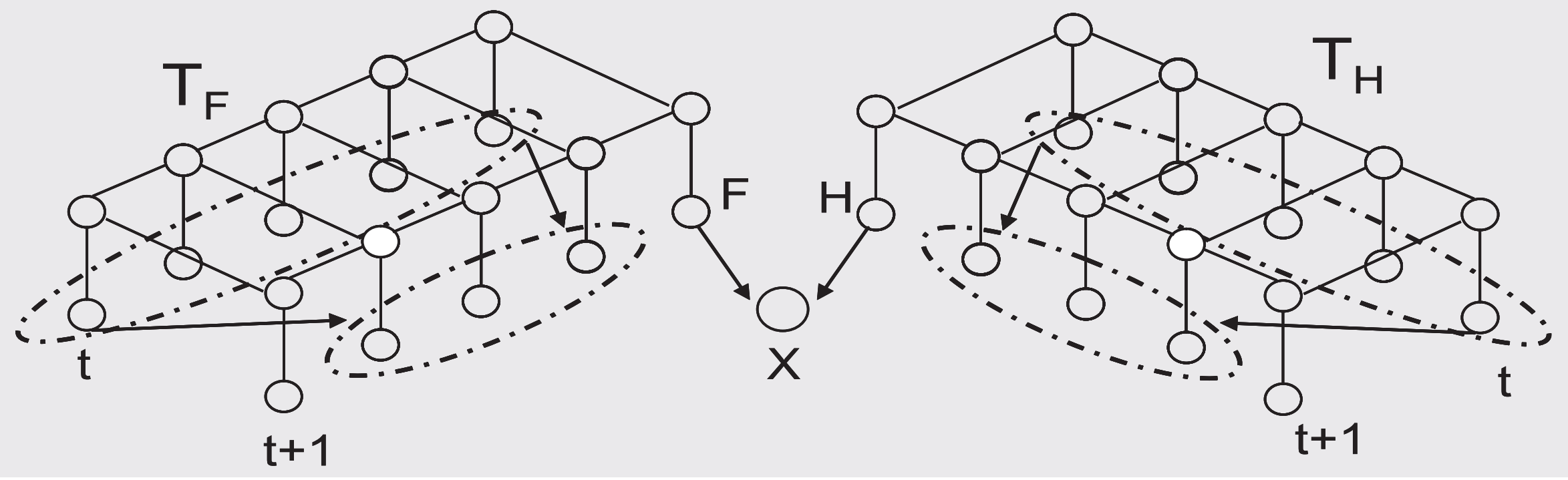
Missing Data.



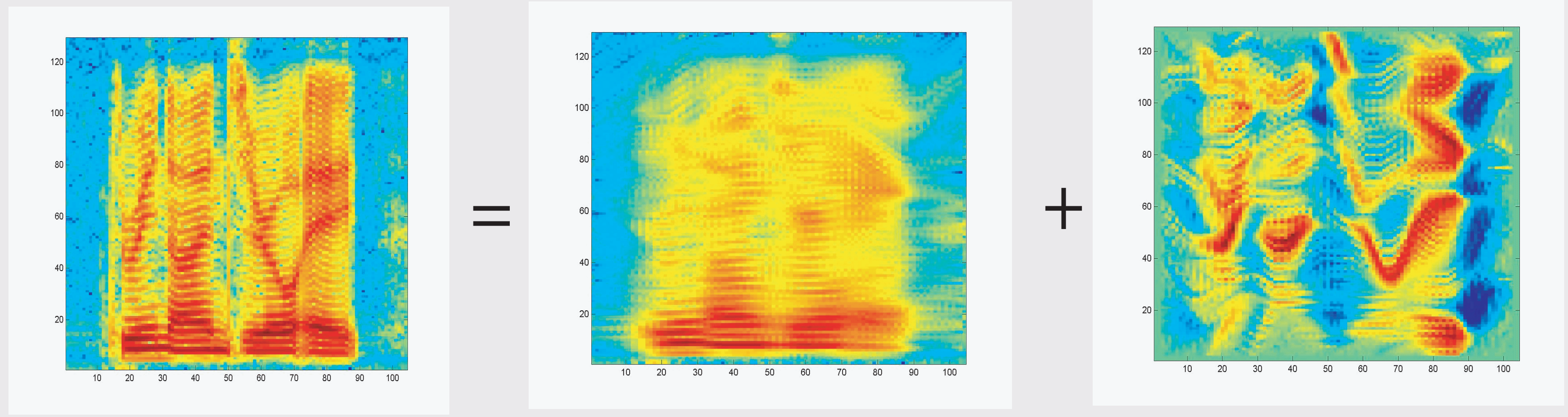
Speech Production Model

$x[n] = h[n] * u[n]$ (Time Domain). $X[\omega] = H[\omega]\, U[\omega]$ (Freq. Domain).

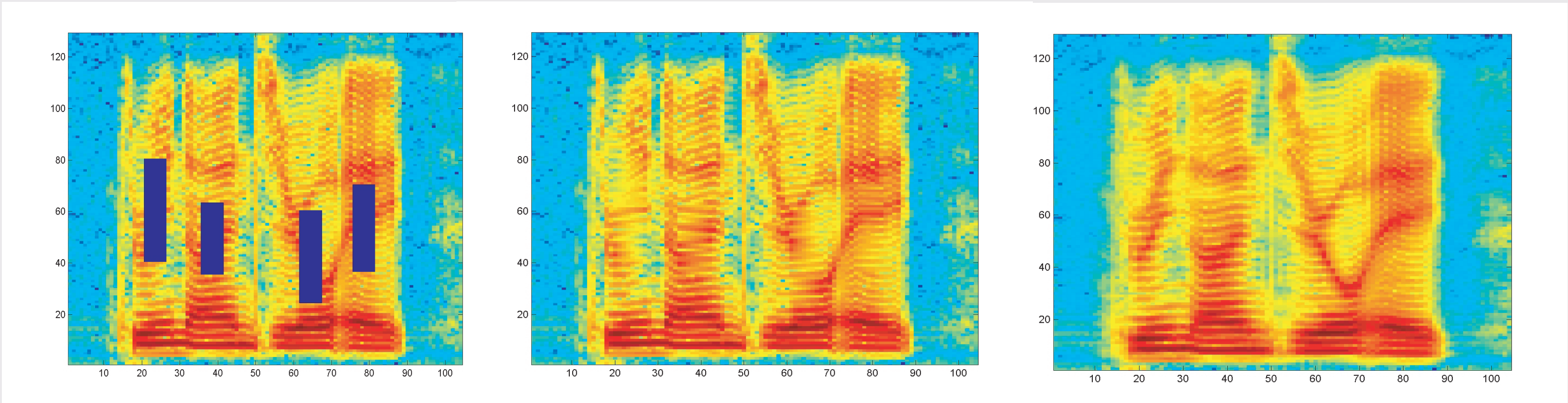
$\log(X[\omega]) = \log(H[\omega]) + \log(U[\omega])$ (Log. Freq. Domain).



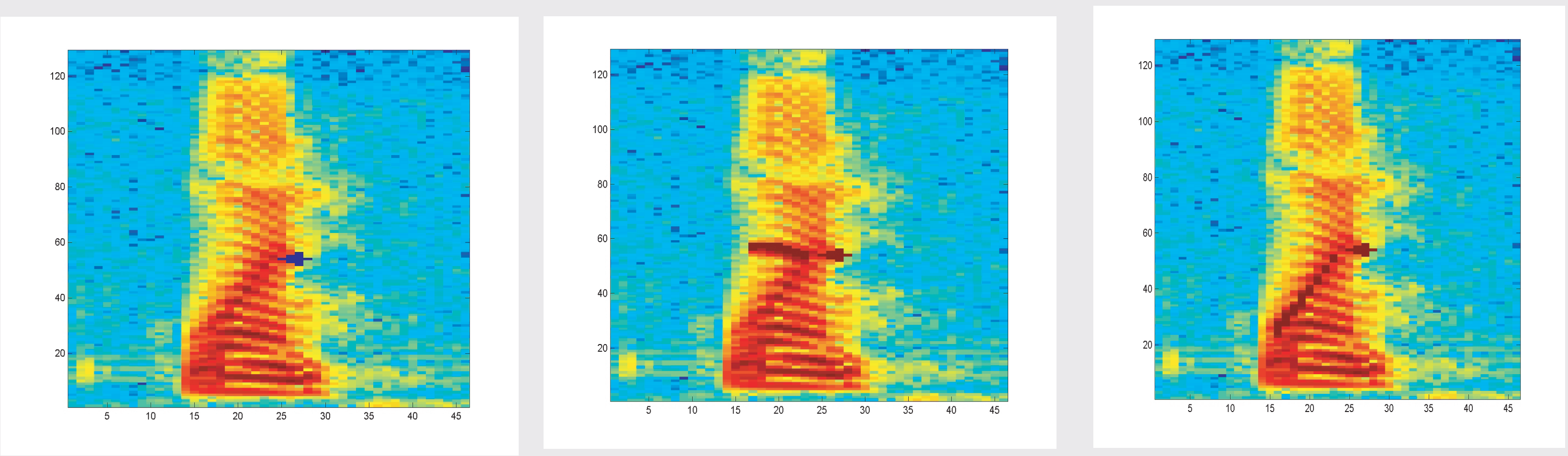
Separation Example



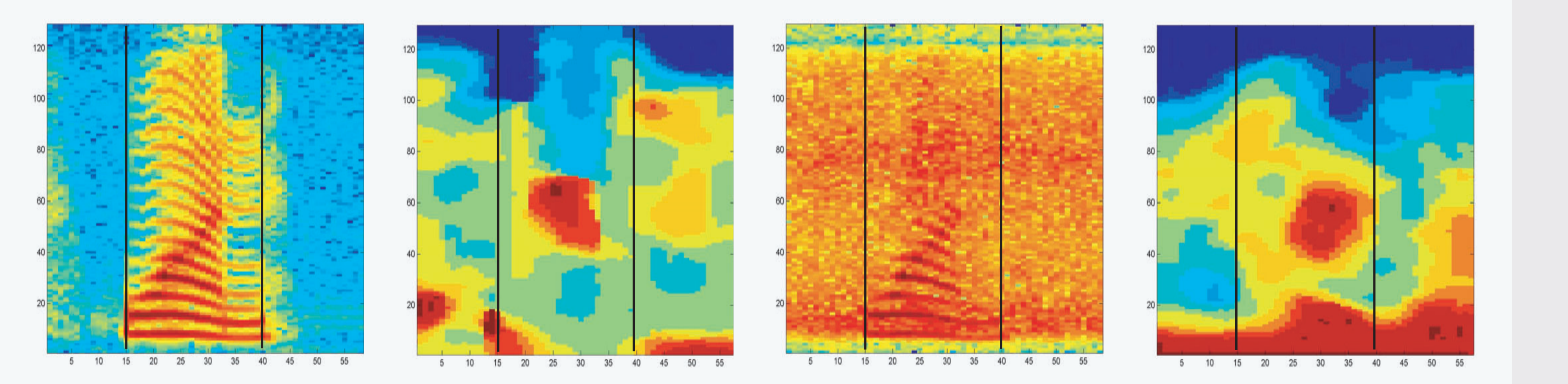
Missing Data.



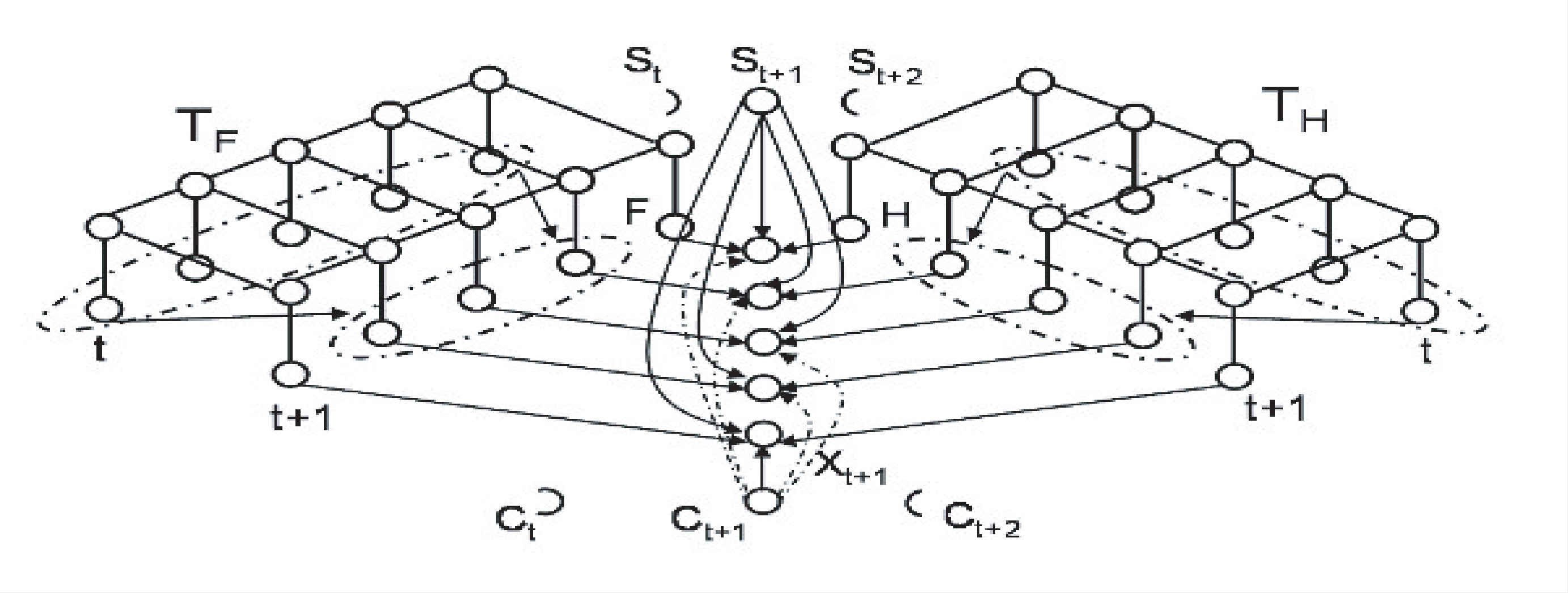
Formants and Harmonics Tracking



Robustness to noise.



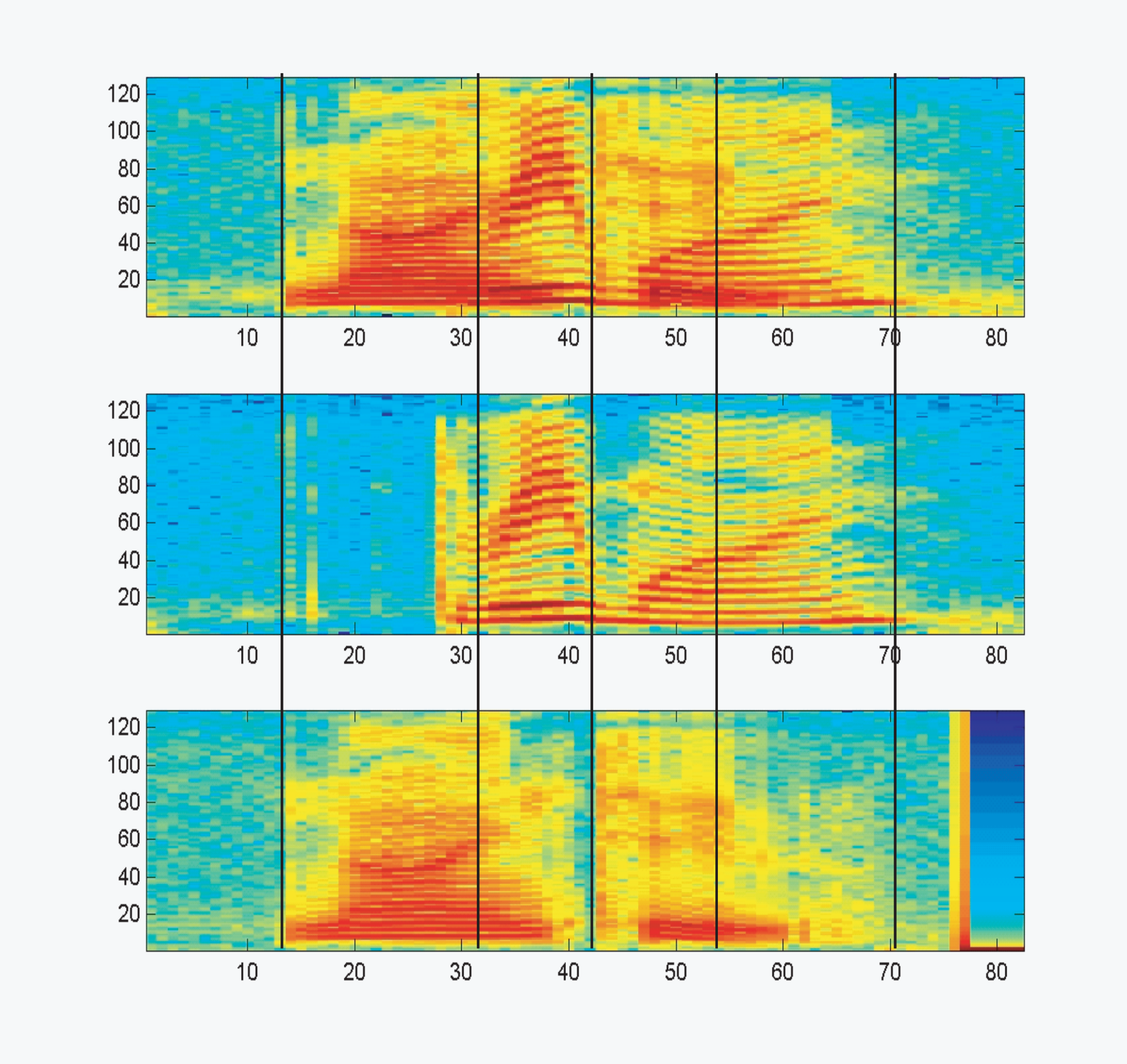
Tracking and and Matching Model



When variable $C_t = 0$; The system tracks a given frequency bin from its context. When $C_t=1$; The systems matches the frequency bin with the correspondent coefficient from one of the states S.

$$p(X_t^k | H_t^k, F_t^k, S_t=j, C_t) = \begin{cases} N(X_t^k; H_t^k + F_t^k, \Delta) & C_t = 0; \\ N(X_t^k; u_j^k, \Sigma_j) & C_t = 1; \end{cases}$$

Example with a mixture of two speakers.



Current work

- Tracking and Matching overlapping patches.
- Identify Regions.
- Cluster Regions and Assign Labels.
- Propagate Labels Using Learned Transformation Maps.
- Learned Speaker Models from Patches and Dissambiguate.