Speaking and Singing Voice Alignment with Deep Neural Networks

Yann Teytaut

Dr. Guillaume Doras

Dr. Axel Roebel

Analysis/Synthesis – IRCAM (STMS UMR 9912) – Paris, France

ircam

cnrs

SORBONNE UNIVERSITÉ

MINISTÈRE DE LA CULTURE
Liberté
Égalité
Fraternité

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

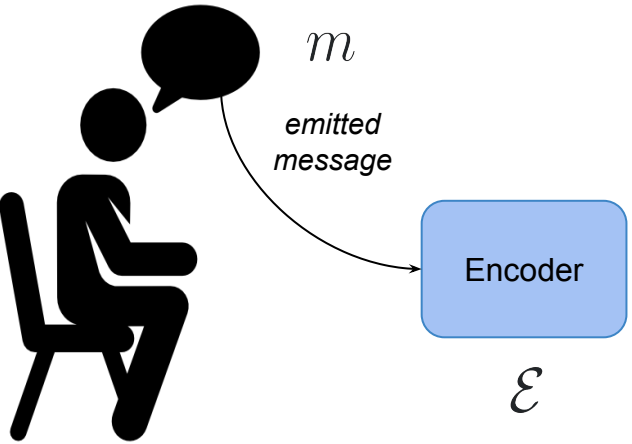
. Human communication

. Human communication

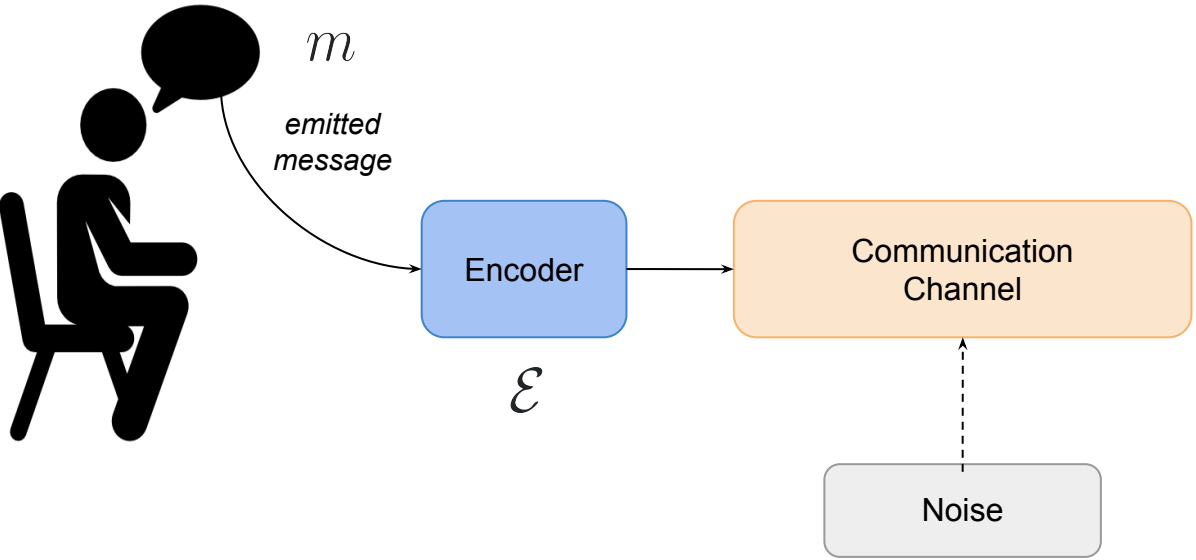
m

emitted message

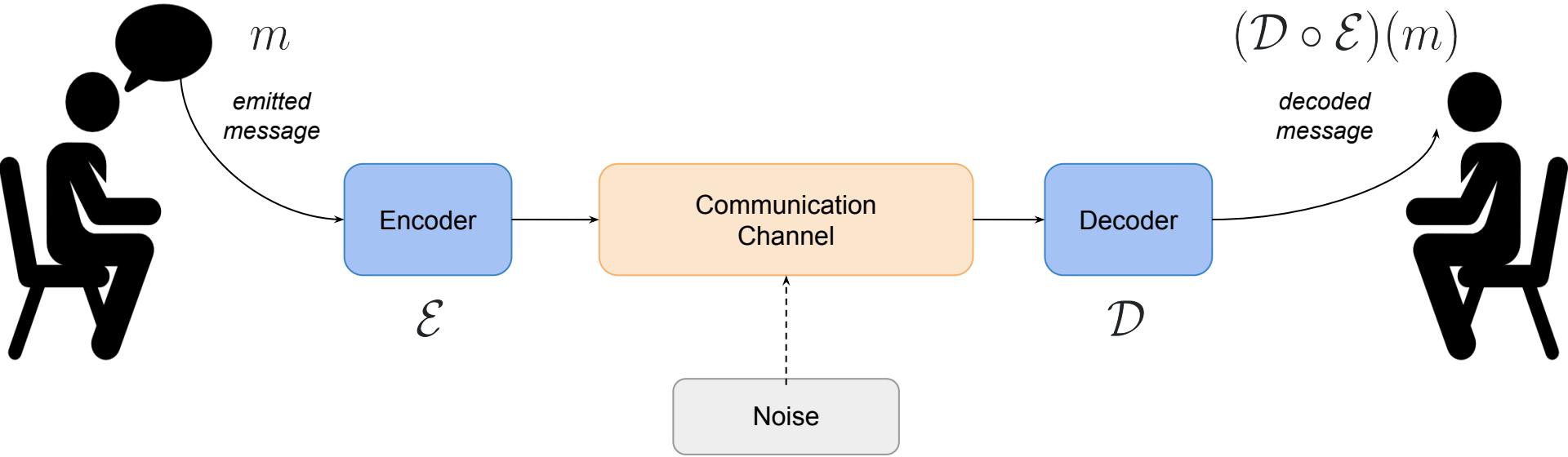
. Human communication



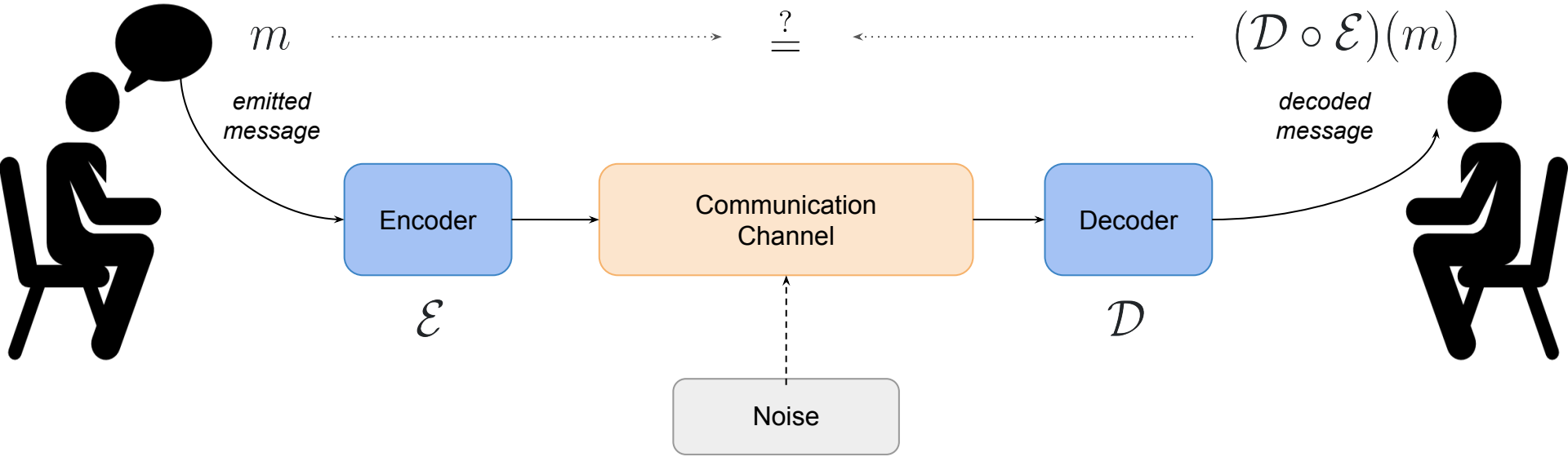
. Human communication



. Human communication



. Human communication



. Context – Temporality

. Context – Temporality

. Context – Temporality

Responding

Listening

Following

Adapting

Coordinating

Synchronizing

. Context – Temporality

Responding

Listening

Following

Adapting

Coordinating

Synchronizing

Aligning

. Motivations

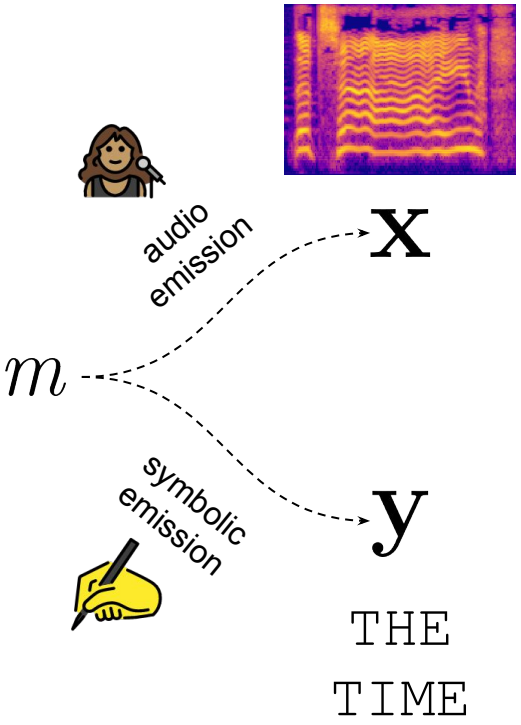
Uncovering the *temporal relationships* in audio and voice signals is of utmost importance for their *analysis*

When is a note played?
When is a word pronounced?
When is there a structural change in the audio?
.
.
.

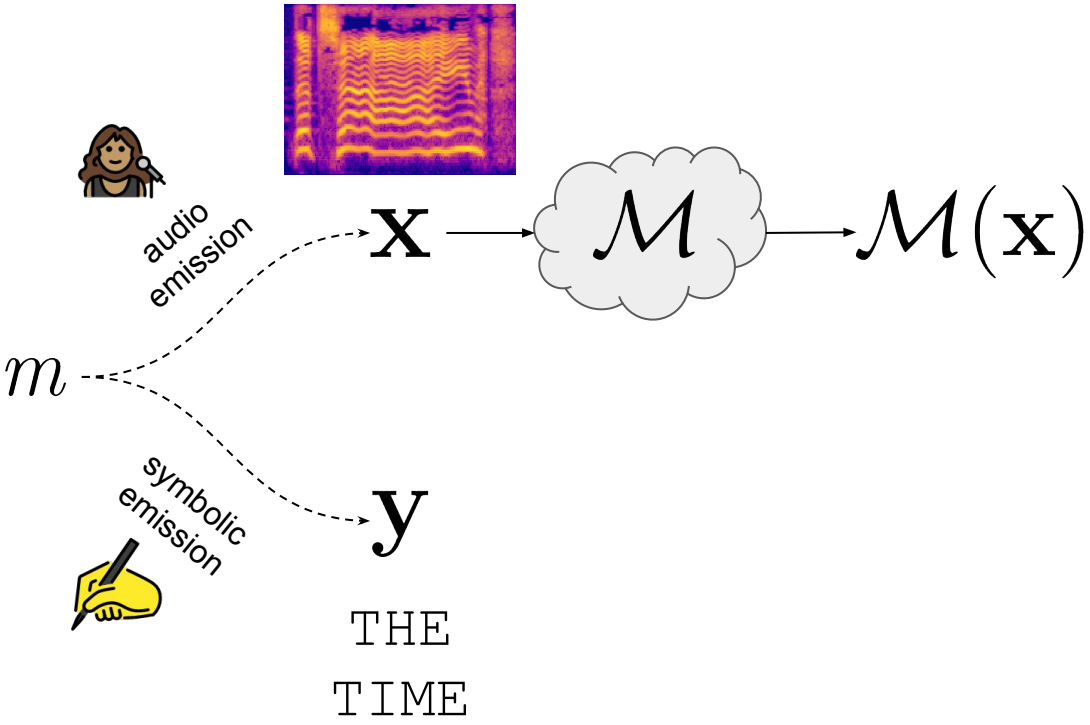
. Voice alignment

m

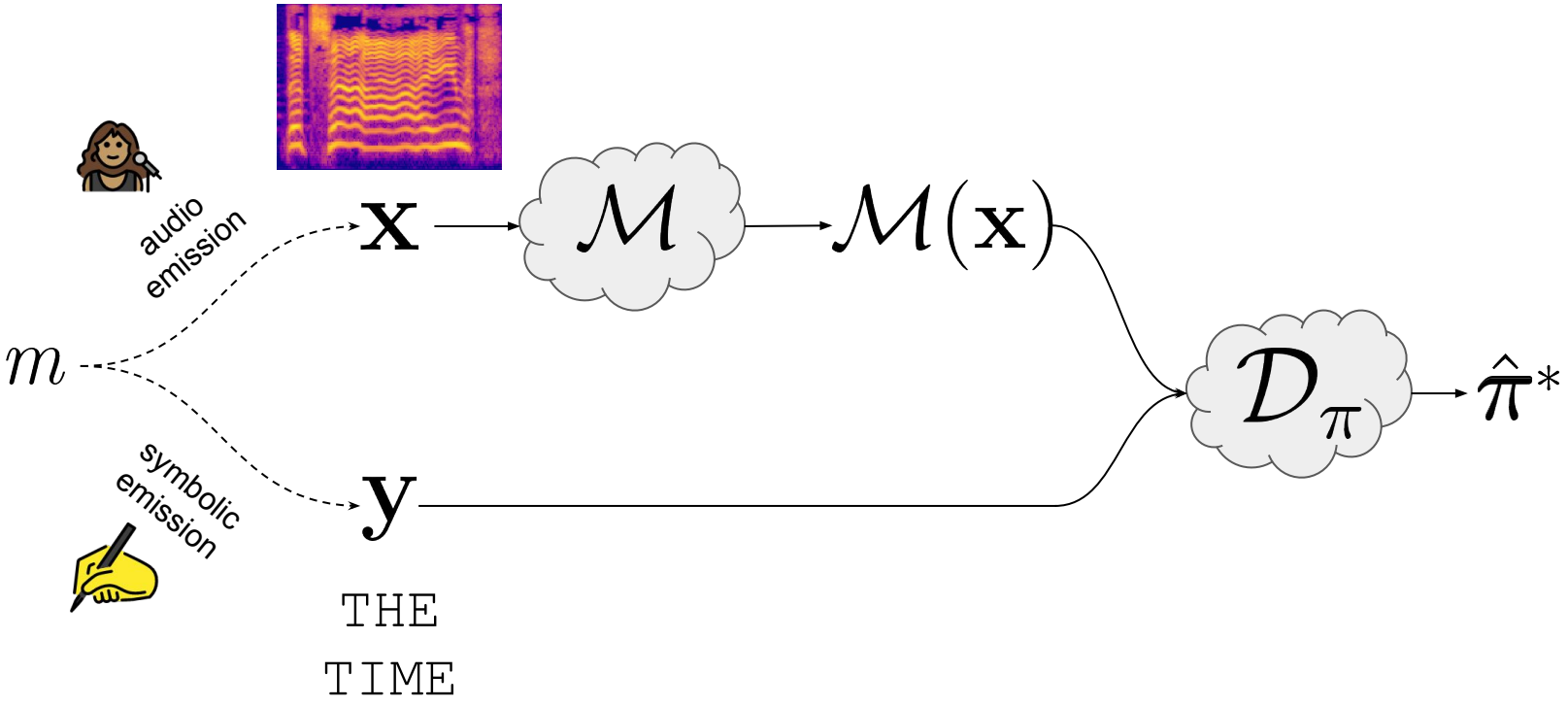
. Voice alignment



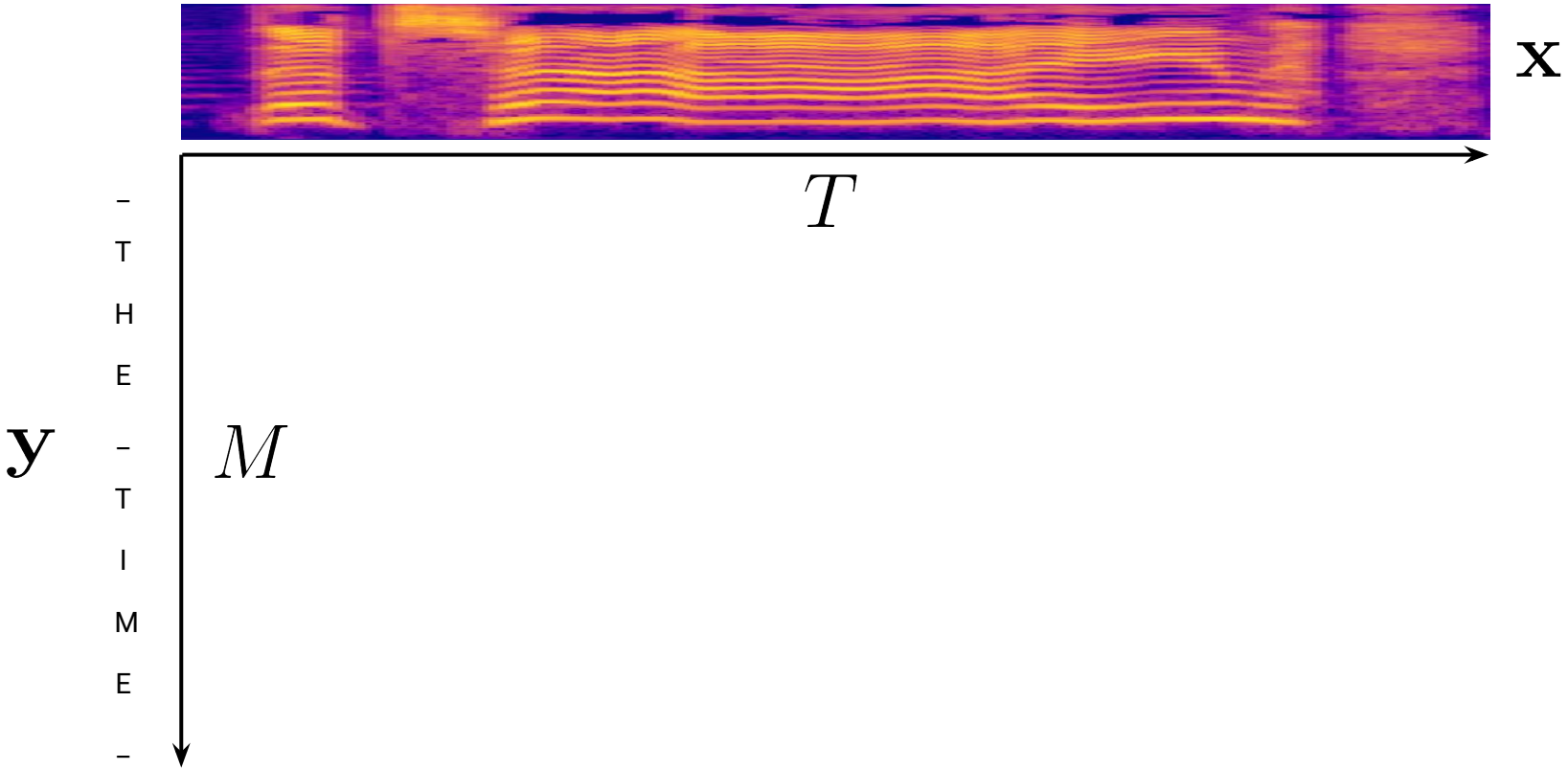
. Voice alignment



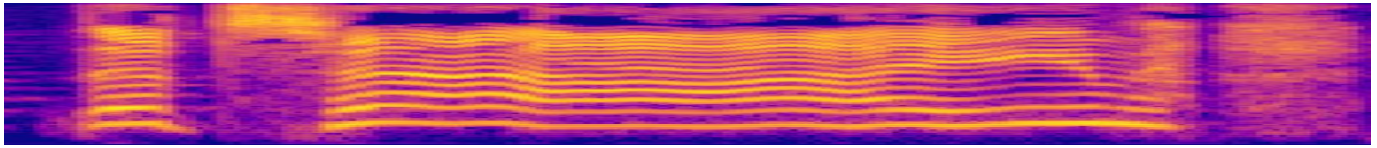
. Voice alignment



. Optimal pathway



. Optimal pathway



x

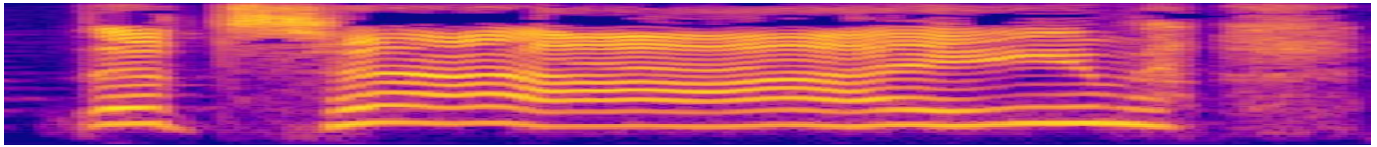
y

– T H E – T I M E –

M

T

. Optimal pathway



x

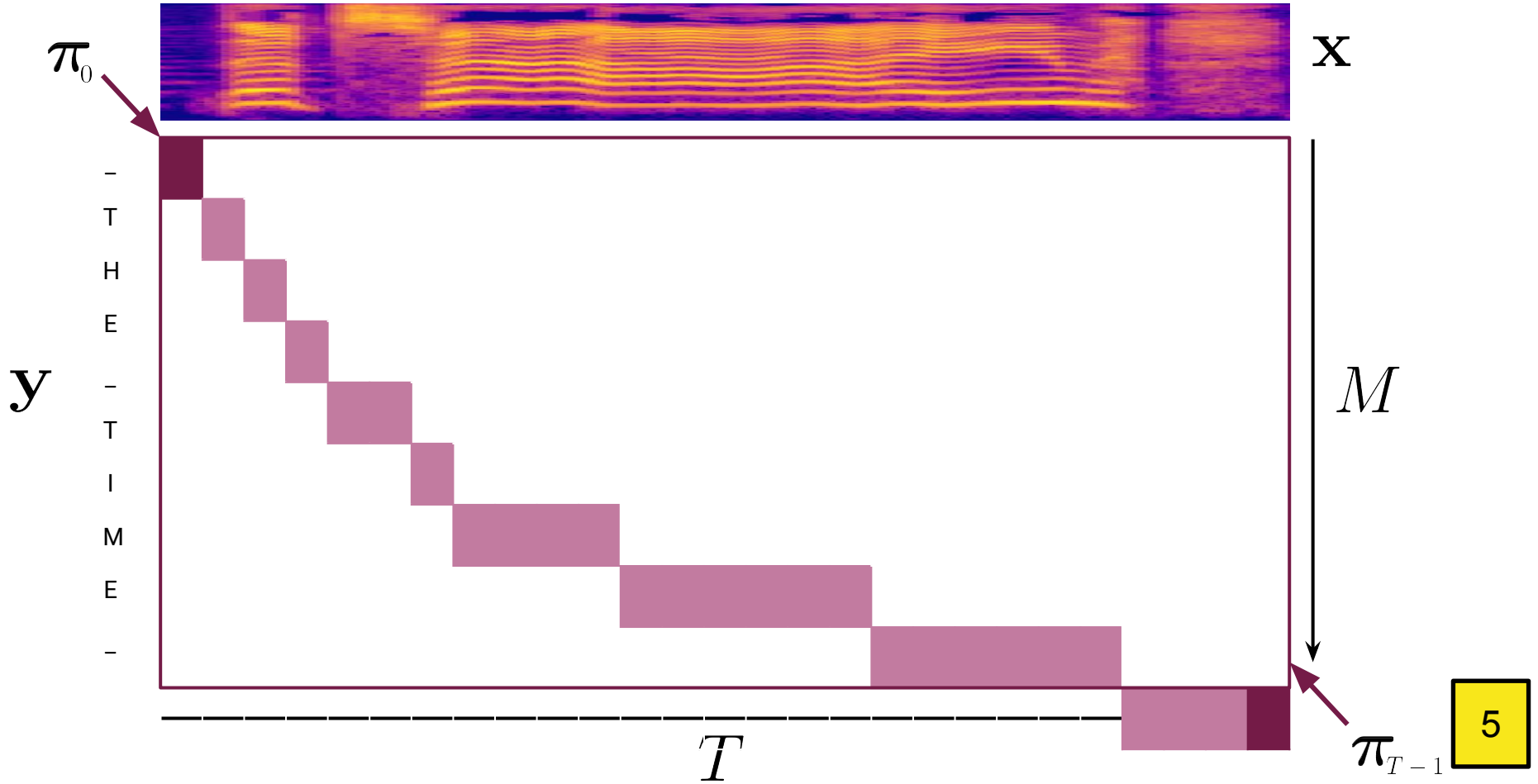
y

– T H E – T I M E –

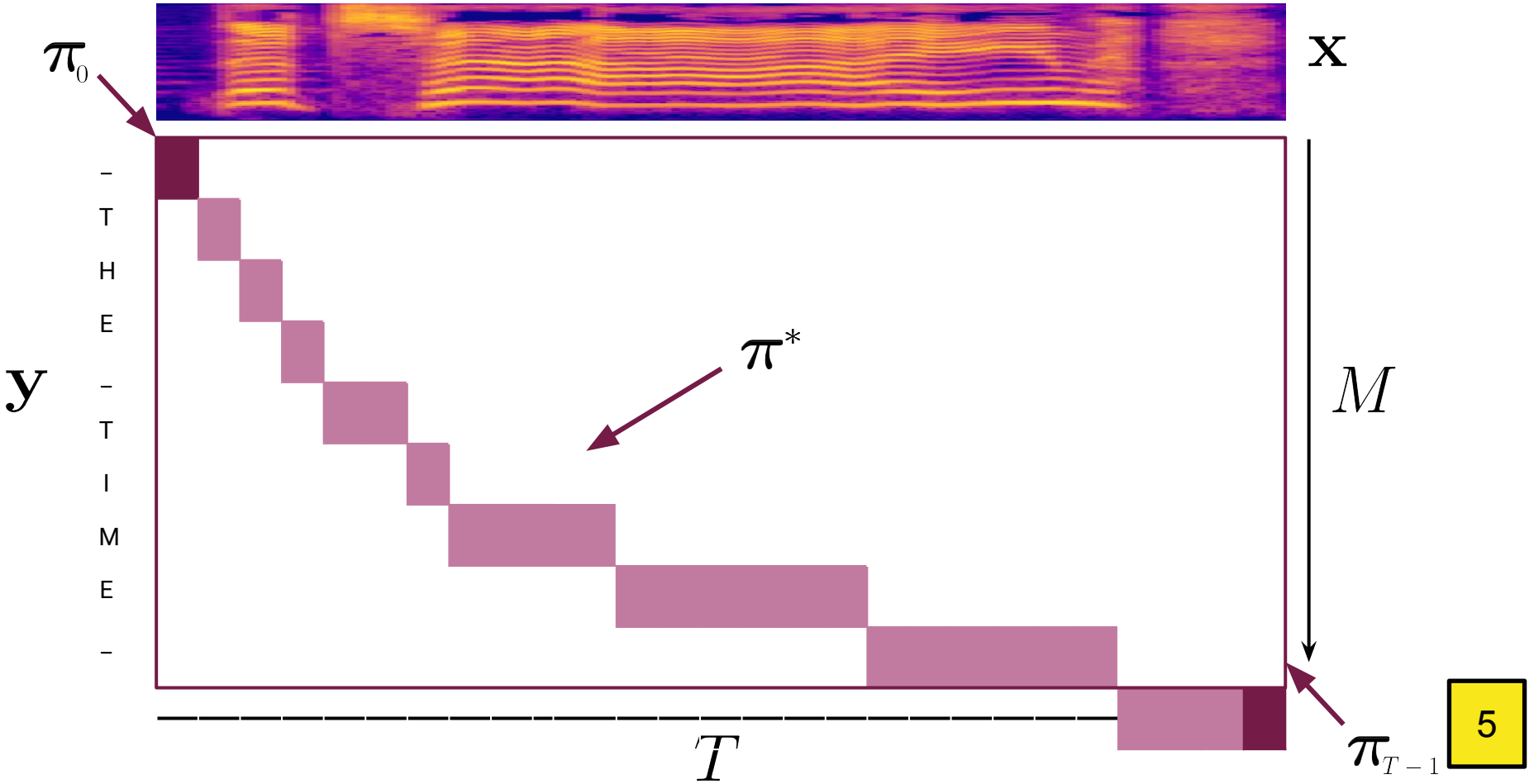
M

T

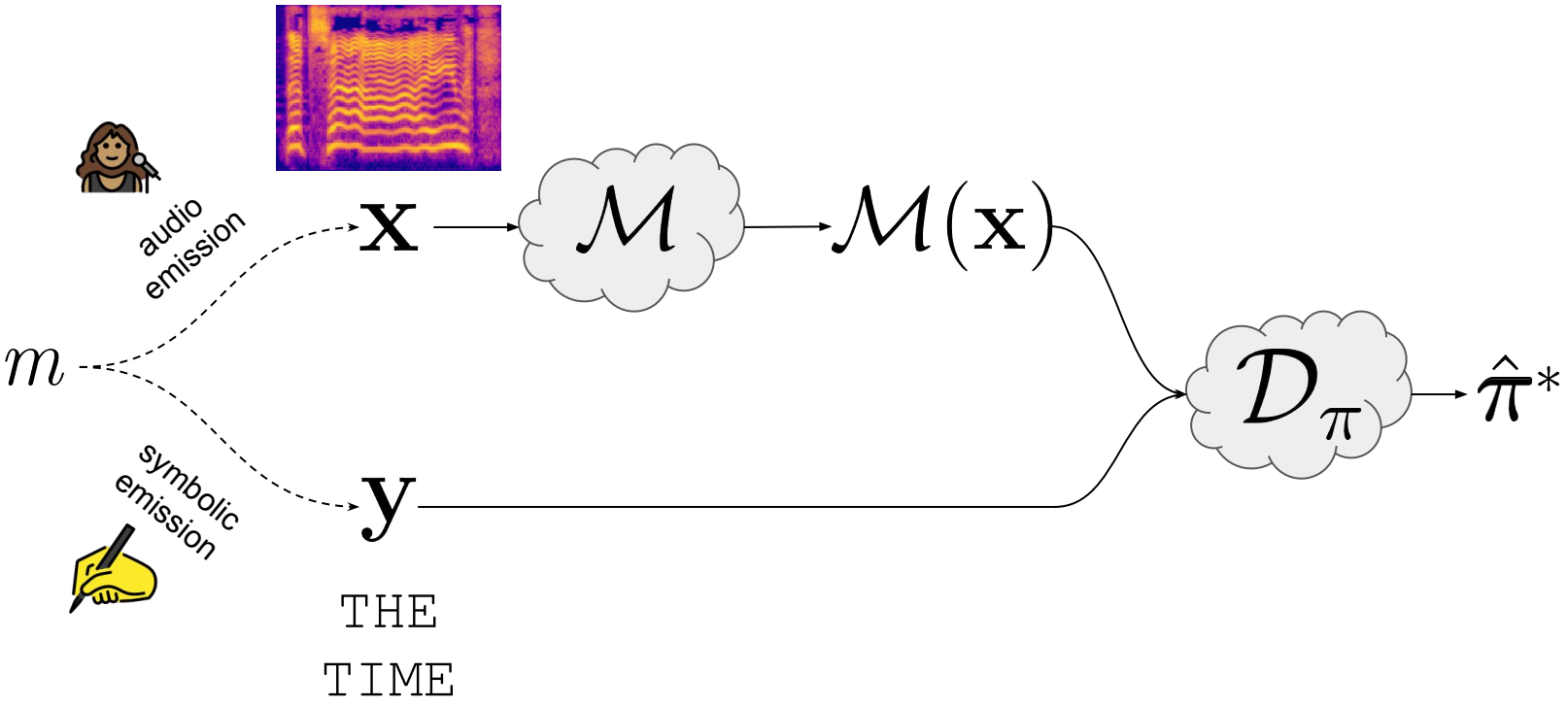
. Optimal pathway



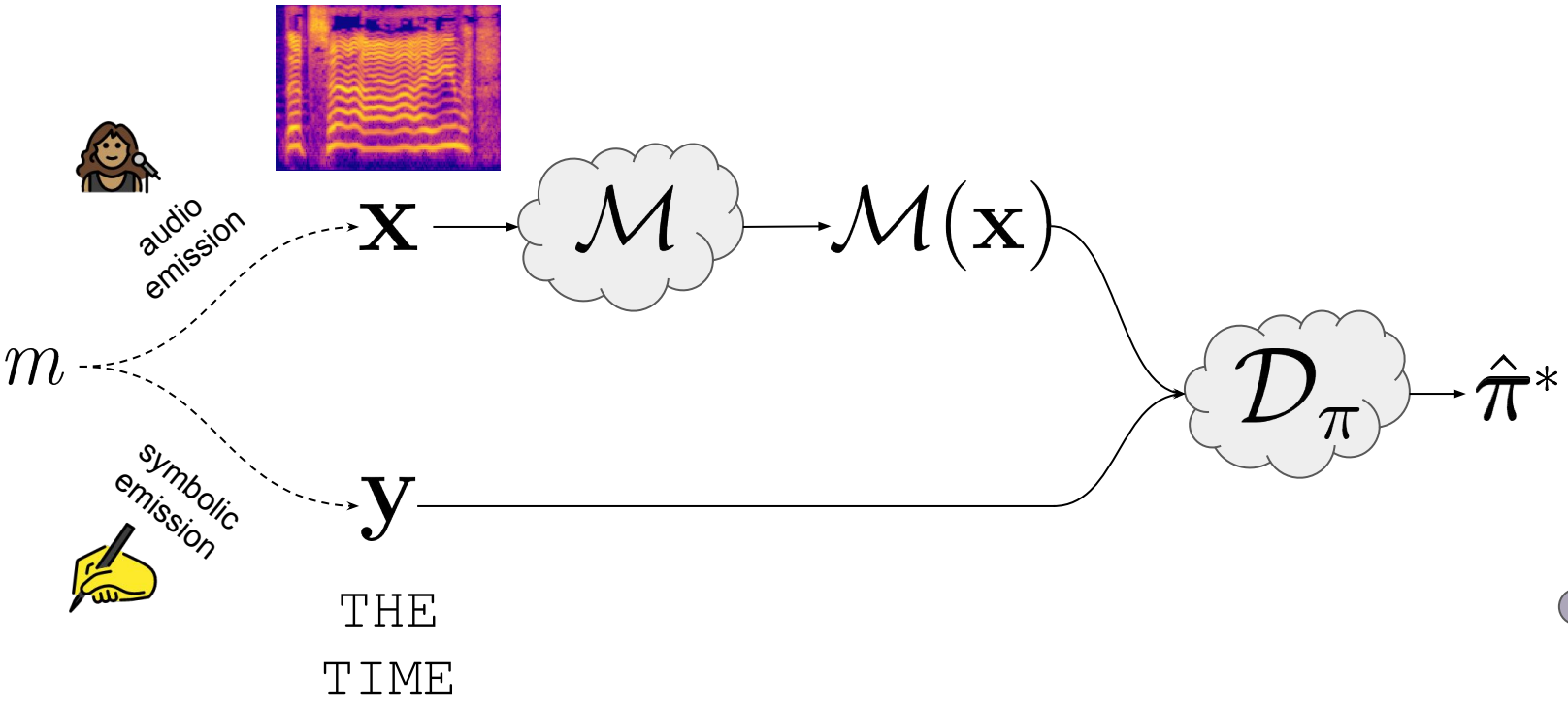
. Optimal pathway



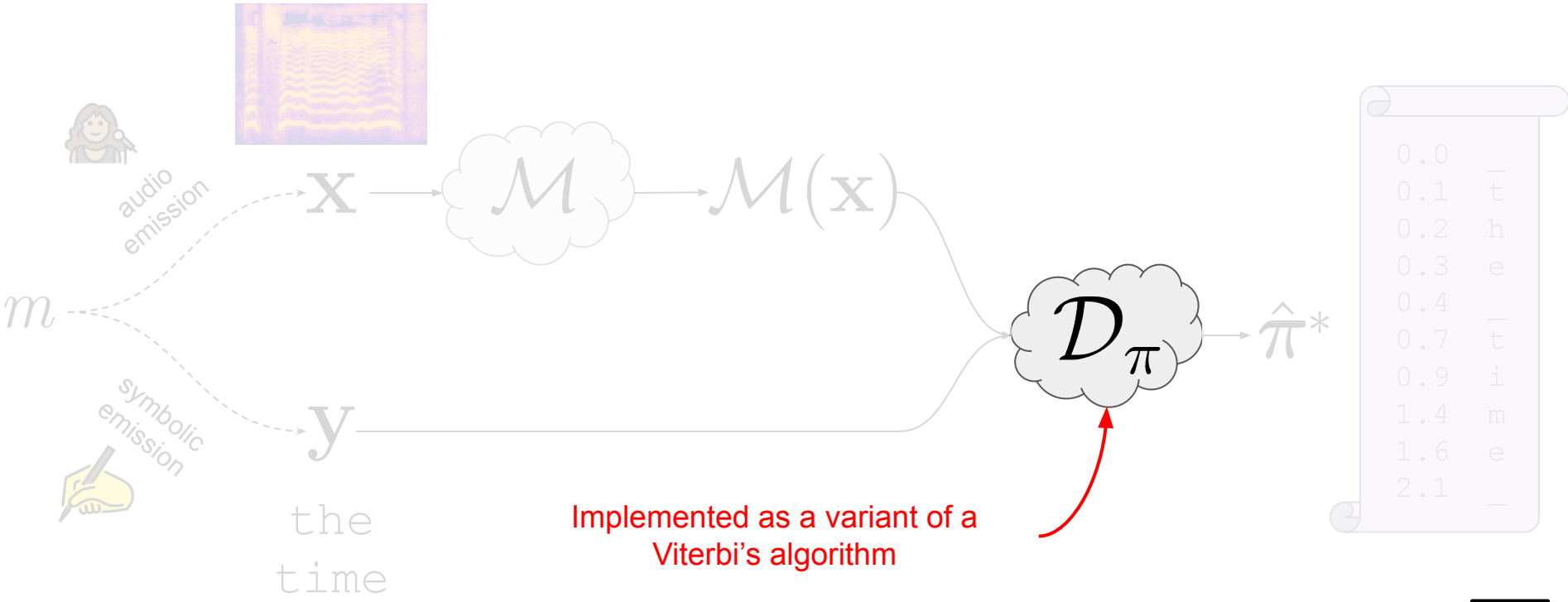
. Overview


audio emission
symbolic emission
m
$\mathbf{x}$
$\mathbf{y}$
$\mathcal{M}$
$\mathcal{M}(\mathbf{x})$
$\mathcal{D}_\pi$
$\hat{\pi}^*$
THE
TIME

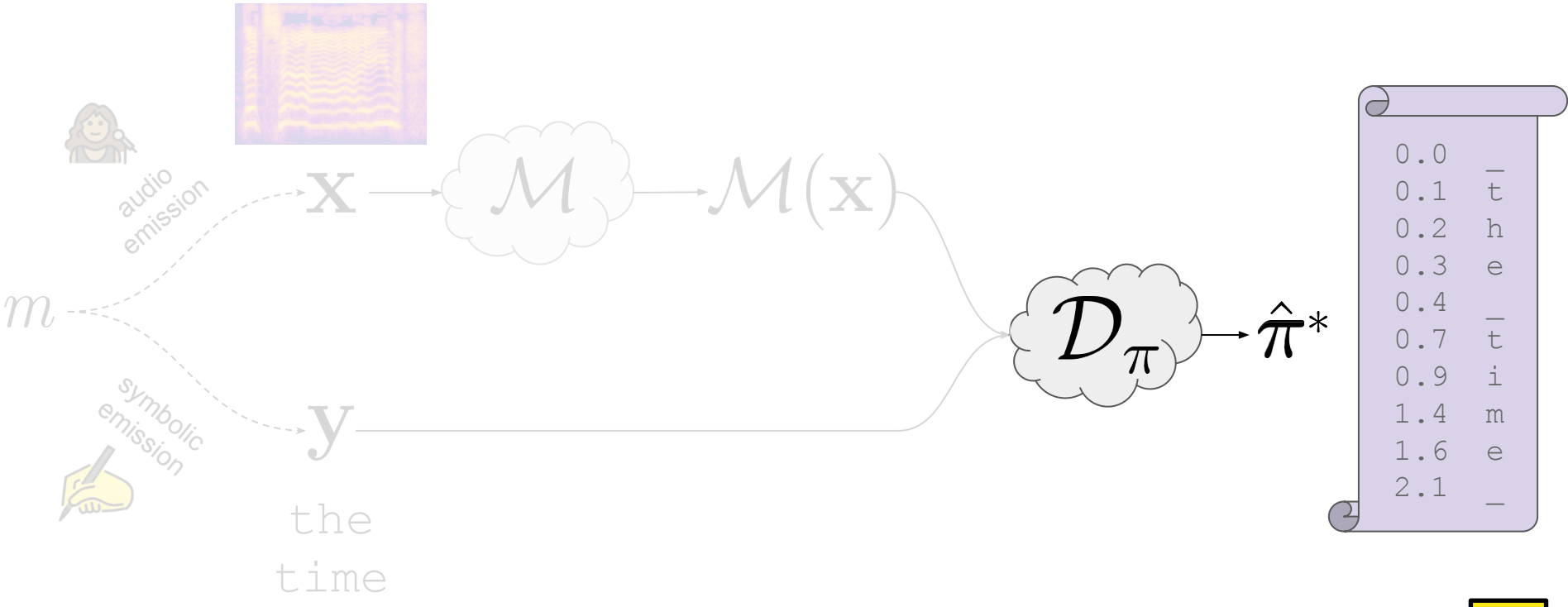
. Overview


audio emission
symbolic emission
m
$\mathbf{x}$
$\mathbf{y}$
$\mathcal{M}$
$\mathcal{M}(\mathbf{x})$
$\mathcal{D}_\pi$
$\hat{\pi}^*$
THE
TIME
0.0 _
0.1 T
0.2 H
0.3 E
0.4 _
0.7 T
0.9 I
1.4 M
1.6 E
2.1 _

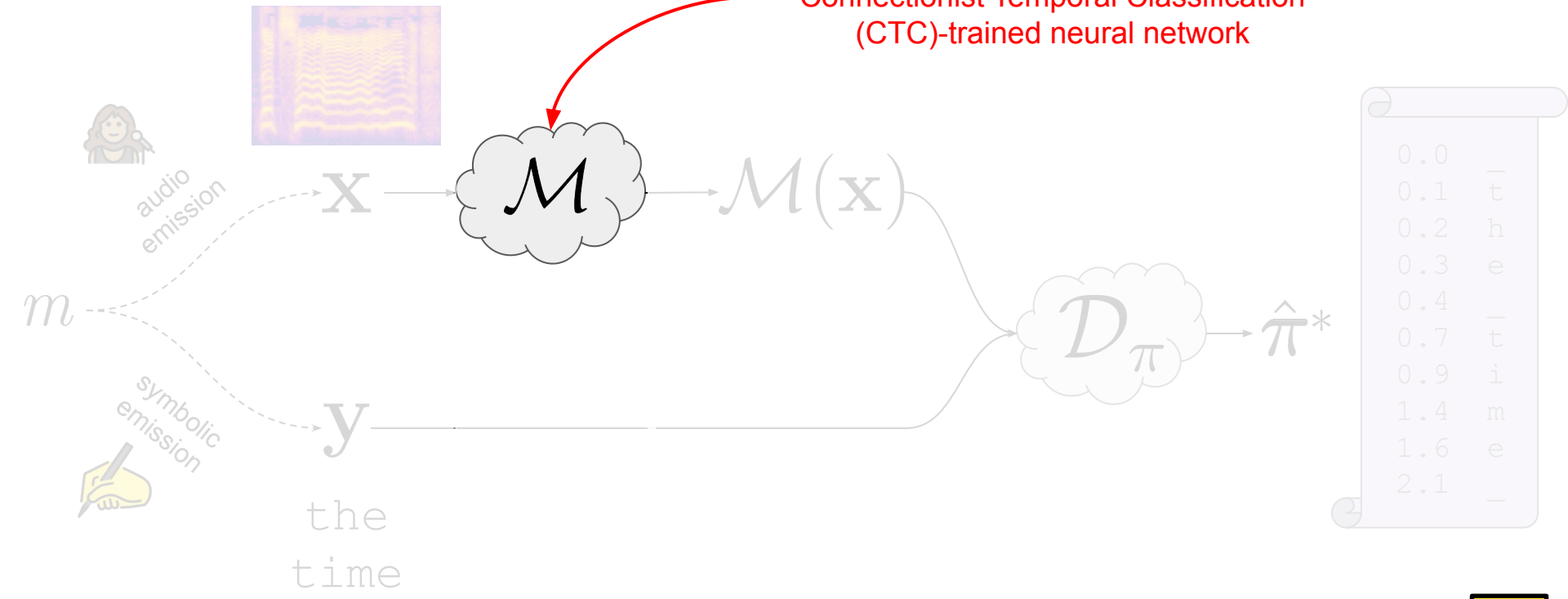
. Path decoding



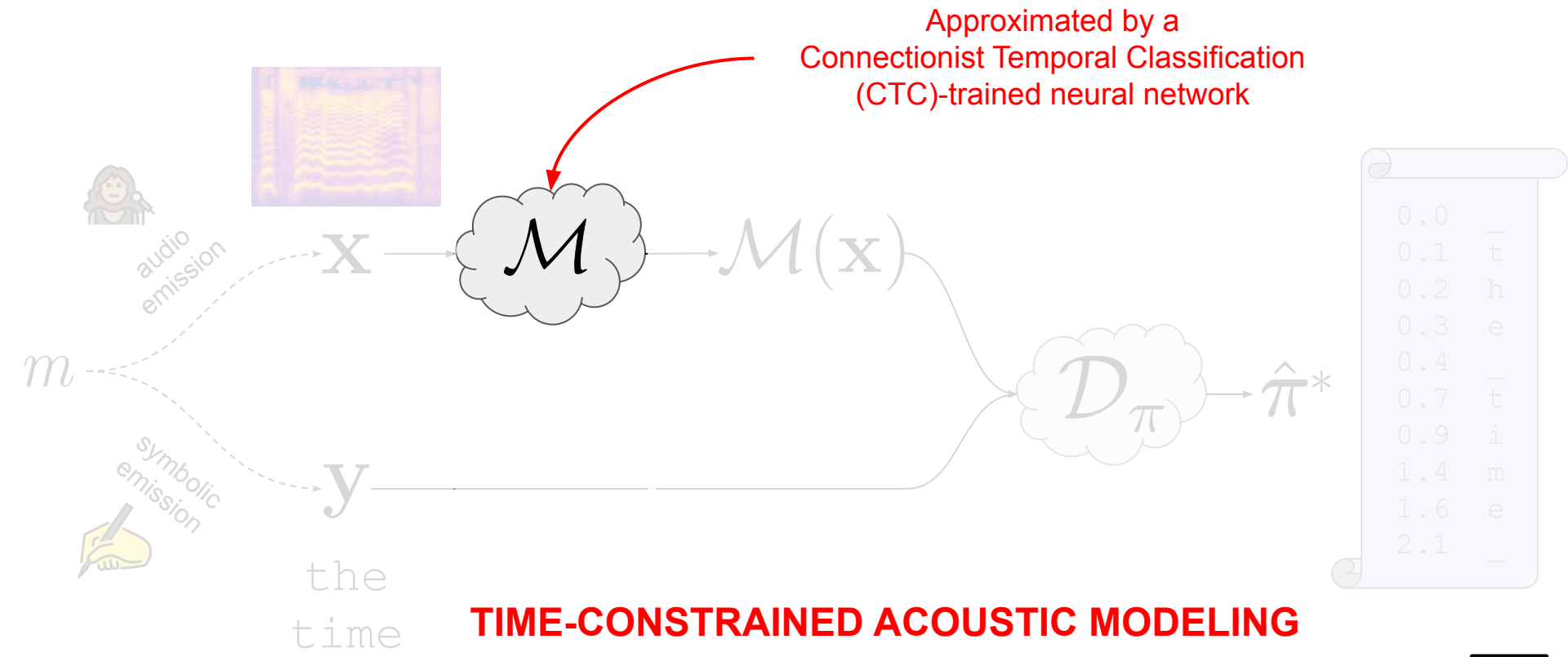
. Path decoding



. Acoustic modeling



. Acoustic modeling



TIME-CONSTRAINED ACOUSTIC MODELING

. Acoustic modeling

Approximated by a
Connectionist Temporal Classification
(CTC)-trained neural network

audio emission

$\mathbf{x}$

$\mathcal{M}$

$\mathcal{M}(\mathbf{x})$

m

$\mathcal{D}_\pi$

$\hat{\pi}^*$

symbolic emission

$\mathbf{y}$

the time

0.0 _
0.1 t
0.2 h
0.3 e
0.4 _
0.7 t
0.9 i
1.4 m
1.6 e
2.1 _

TIME-CONSTRAINED ACOUSTIC MODELING
DETAILS ON POSTER :)