Multimodal Evolution Inference
Temporal probabilistic models in multimodal perception

Xavier Alameda-Pineda

January 20, 2014

Abstract

We derive an EM for a big graphical model.

1 Notation

1.1 Generics

Scalar, vector and matrix Italic means scalar, italic bold means vector and bold means matrix. For example, in the same order:

$$A \qquad \boldsymbol{A} \qquad \mathbf{A}$$

Collection of indexed variables Whenever a variable has an index, such as A_n the following notations hold:

$A_{1:N}$ is the set of A_n for all values of n from 1 to N.

$A_{1:m:N}$ is the set of A_n for all values of n from 1 to N except fro m.

Vector and matrix components Whenever $\boldsymbol{A}$ is a vector $\boldsymbol{A}[n]$ denotes the n-th component of the vector. The same intuition holds for $\mathbf{A}[i,j]$ when $\mathbf{A}$ is a matrix.

Important sets We need to consider different sets of matrices and vectors to simplify the notations:

- $\mathcal{CM}^D \subset \mathbb{R}^{D\times D}$ is the set of covariance matrices of dimension D. That is $\mathbf{x} \in \mathcal{CM}^D$ if and only if $\mathbf{x}$ is symmetric and positive definite.
- $\mathcal{SV}^D \subset \mathbb{R}^D$ is the set of stochastic vectors of dimension D. That is $\boldsymbol{x} \in \mathcal{SV}^D$ if and only if $\boldsymbol{x}[d] \geq 0 \, \forall d = 1, \ldots, D$ and $\sum_{d=1}^D \boldsymbol{x}[d] = 1$.
- $\mathcal{SM}^{D\times K} \subset \mathbb{R}^{D\times K}$ is the set of row-stochastic matrices of dimension $D\times K$. That is $\mathbf{x} \in \mathcal{SV}^{D\times K}$ if and only if $\mathbf{x}[d,k] \geq 0 \, \forall d = 1, \ldots, D, k = 1, \ldots, K$ and $\sum_{k=1}^K \mathbf{x}[d,k] = 1$, $\forall d = 1, \ldots, D$.

Finally, we will use $\mathfrak{D}(\cdot)$ to denote the dimensionality of the model.

1.2 In the model

N Number of sources.

1.2.1 Source Position

- D dimension of the source space, $\mathbb{R}^D$.
- $\boldsymbol{S}_{t,n}$ position of the n^{th} source at time t. $\boldsymbol{S}_t$ concatenation of the N position vectors.
- $\boldsymbol{\nu}_n$, $\boldsymbol{\Omega}_n$, parameters of the Gaussian for $\boldsymbol{S}_{1,n}$, $\boldsymbol{\nu}_n \in \mathbb{R}^D$ and $\boldsymbol{\Omega}_n \in \mathcal{CM}^D$.
- $\boldsymbol{\Lambda}_{n,0}$, $\boldsymbol{\Lambda}_{n,1}$, parameters of the diffusion dynamics (not moving and moving). $\boldsymbol{\Lambda}_{n,0}, \boldsymbol{\Lambda}_{n,1} \in \mathcal{CM}^D$.
- $\theta^S = \{\boldsymbol{\nu}_n, \boldsymbol{\Omega}_n, \boldsymbol{\Lambda}_{n,0}, \boldsymbol{\Lambda}_{n,1}\}_{n=1}^N$.

1.2.2 Auditory status

- $\boldsymbol{A}_t \in \{0,1\}^N$ (or $\boldsymbol{A}_t \in \{1, \ldots, 2^N\}$), auditory status of the N sources at time t.
- $\mathbf{T}^A, \boldsymbol{\lambda}^A$, transition matrix and prior vector for $\boldsymbol{A}_t$.
- $\mathbf{T}^A \in \mathcal{SM}^{2^N\times 2^N}$ and $\boldsymbol{\lambda}^A \in \mathcal{SV}^{2^N}$.
- $\theta^A = \{\mathbf{T}^A, \boldsymbol{\lambda}^A\}$.

1.2.3 Visibility and Motion status

The follow the same logic, changing F by M.

- $F_{t,n}$ is the visibility of the n^{th} source at time t.
- $\mathbf{T}_n^F, \boldsymbol{\lambda}_n^F$, transition matrix and prior vector for $F_{t,n}$.
- $\mathbf{T}_n^F \in \mathcal{SM}^{2\times 2}$ and $\boldsymbol{\lambda}_n^F \in \mathcal{SV}^2$.
- $\theta^F = \{\mathbf{T}_n^F, \boldsymbol{\lambda}_n^F\}_{n=1}^N$.

1.2.4 Mel Frequency Cepstral Coefficients [MFCC]

D_V dimension of the MFCC, $\boldsymbol{V} \in \mathbb{R}^{D_V}$.

- G_V number of components of the GMM.
- K_t^V number of MFCC observations at time t.
- $Z_{t,k}^V \in \{1, \ldots, G_V\}$ observation-to-component assignment variable of the k^{th} observation at time t.
- $\boldsymbol{V}_{t,k}$ k^{th} MFCC observation at time t.
- There is one GMM per auditory status $1, \ldots, 2^N$.

- $\boldsymbol{\pi}_j^V$ prior vector of the mixing coefficients, $j = 1, \dots, 2^N$.
- $\boldsymbol{\mu}_{j,g}^V, \boldsymbol{\Sigma}_{j,g}^V$ mean and covariance matrix of the g^{th} component of the j^{th} GMM, $g = 1, \dots, G_V$, $j = 1, \dots, 2^N$.
- $\theta^{ZV} = \{\boldsymbol{\pi}_1^V, \dots, \boldsymbol{\pi}_{2^N}^V\}$, $\boldsymbol{\pi}_j^V \in \mathcal{SV}^{G_V}$.
- $\theta^V = \{\boldsymbol{\mu}_{j,g}^V, \boldsymbol{\Sigma}_{j,g}^V\}_{g=1,j=1}^{G_V,2^N}$, $\boldsymbol{\mu}_{j,g}^V \in \mathbb{R}^{G_V}$, $\boldsymbol{\Sigma}_{j,g}^V \in \mathcal{CM}^{D_V}$.

1.2.5 Spectrogram [SPEC]

- $\mathbf{P}^L, \mathbf{P}^R$ left and right spectrograms. $\mathbf{P}^L, \mathbf{P}^R \in \mathbb{C}^{P_F \times P_I}$.
- G_P number of components generating the spectrograms.
- $\mathbf{C}_{t,g}$, g^{th} component at time t.
- $\boldsymbol{\omega}_g, \boldsymbol{\xi}_g$, spectral pattern and temporal activation pattern of the g^{th} component.
- $\boldsymbol{\Xi}^L, \boldsymbol{\Xi}^R$ left and right mixing matrices. $\boldsymbol{\Xi}^L, \boldsymbol{\Xi}^R \in \mathbb{C}^{P_F \times G_P}$.
- $\boldsymbol{\Sigma}_f^{PL}, \boldsymbol{\Sigma}_f^{PR}$ noise variances of the f^{th} frequency bin (L&R).
- $\theta^C = \{\boldsymbol{\omega}_g, \boldsymbol{\xi}_g\}_{g=1}^{G_V}$.
- $\theta^P = \{\boldsymbol{\Xi}^L, \boldsymbol{\Xi}^R, \{\boldsymbol{\Sigma}_f^{PL}, \boldsymbol{\Sigma}_f^{PR}\}_{f=1}^{P_F}\}$.

1.2.6 Binaural SSL [BINAURAL] [MOTION]

Motion and binaural features follow the very same model, replacing B by Y, and conditioning with respect to $\boldsymbol{M}_t$ instead of $\boldsymbol{A}_t$.

- K_t^B number of binaural observations at time t.
- $\boldsymbol{B}_{t,k}$ k^{th} binaural observation at time t.
- $\boldsymbol{B}_{t,k}$ follows a N-component GMM.
- $\boldsymbol{W}_t^B$ mixing priors, following a Dirichlet.
- $\boldsymbol{\alpha}_n^B$, parameters of the n^{th} component of the Dirichlet.
- $Z_{t,k}^B$ observation-to-component (source).
- $\boldsymbol{\Sigma}_{t,n}^B$, covariance matrix of the n^{th}, $\boldsymbol{\Sigma}_{t,n}^B \in \mathcal{CM}^D$.
- $\theta^{WB} = \{\boldsymbol{\alpha}_1^B, \dots, \boldsymbol{\alpha}_N^B\}$.
- $\theta^B = \{\boldsymbol{\Sigma}_{t,n}^B\}_{t=1,n=1}^{T,N}$.

1.2.7 [FACIAL]

- K_t^F number of detected faces at time t.
- e face detector false positive rate.
- $\boldsymbol{Q}_{t,k} \in \mathbb{R}^D$ k^{th} detection position at time t.
- $\boldsymbol{\Sigma}^Q$ covariance matrix (face detector localization error).
- $Z_{t,k}^R$ pose detection (among G_R poses).
- G_F number of facial landmarks.
- $\boldsymbol{R}_{t,k}^G \in \mathbb{R}^{G_F D}$ positions of the landmarks.
- $\boldsymbol{\Sigma}_{n,m}^G$ covariance of the Gaussian describing $\boldsymbol{R}^G$ for the m^{th} pose of the n^{th} source, $\boldsymbol{\Sigma}_{n,m}^G \in \mathcal{CM}^{G_F D}\, \forall n, m$
- D_A dimension of the landmark descriptors.
- $\boldsymbol{R}_{t,k,g}^A \in \mathbb{R}^{D_A}$ descriptor of the g^{th} landmark.
- $\boldsymbol{\mu}_{n,m,g}^A, \boldsymbol{\Sigma}_{n,m,g}^A$ mean and covariance of the Gaussian describing the g^{th} landmark of the m^{th} pose of the n^{th} source.
- G_L number of lip landmarks.
- $\boldsymbol{L}_{t,k,g} \in \mathbb{R}^{D_L}$ lip movement descriptor of the g^{th} landmark.
- $\boldsymbol{\mu}_{n,m,g}^L, \boldsymbol{\Sigma}_{n,m,g}^L$ mean and covariance of the Gaussian describing the g^{th} lip landmark of the m^{th} pose of the n^{th} source.

2 The Model

We present in this section the graphical model, that is the probability description of the relationship between the observed variables and the hidden variables and how this relationship is parametrized. The full model is shown in Figure 1. In the following, we describe the variables and its relations little by little.

We assume the existence of N sources (potential speakers).

2.1 Dynamics

2.1.1 Positions

The N sources are placed in $\mathbb{R}^D$. $\boldsymbol{S}_{t,n} \in \mathbb{R}^D$ denotes the position of the n^{th} source at time t, and $\boldsymbol{S}_t$ is the concatenation of the N vectors, so $\boldsymbol{S}_t \in \mathbb{R}^{ND}$. The dynamics of the positions are modelled separately:

$$\boldsymbol{S}_{t,n} | \boldsymbol{S}_{t-1,n}, M_{t-1,n} \sim \mathcal{N}(\boldsymbol{S}_{t,n}; \boldsymbol{S}_{t-1,n}, \boldsymbol{\Lambda}_{n,M_{t-1,n}}) \tag{2.1}$$

$$\boldsymbol{S}_{1,n} \sim \mathcal{N}(\boldsymbol{S}_{1,n}; \boldsymbol{\nu}_n, \boldsymbol{\Omega}_n). \tag{2.2}$$

This means that for each source there are two possible dynamics, either moving $M_{t-1,n} = 1$ or not moving $M_{t-1,n} = 0$ (see Section 2.1.4). Therefore, we consider two transition matrices per source: dynamic $\boldsymbol{\Lambda}_{n,1}$ and static $\boldsymbol{\Lambda}_{n,0}$. This holds for all time steps, except for the first one, in which is modelled as a multivariate Gaussian with mean $\boldsymbol{\nu}_n$ and covariance $\boldsymbol{\Omega}_n$.

The parameters are: $\theta^S = \{\boldsymbol{\nu}_n, \boldsymbol{\Omega}_n, \boldsymbol{\Lambda}_{n,1}, \boldsymbol{\Lambda}_{n,0}\}_{n=1}^N$, with $\boldsymbol{\nu}_n \in \mathbb{R}^D$, $\boldsymbol{\Lambda}_{n,0:1}, \boldsymbol{\Omega}_n \in \mathcal{CM}^N$ for all n. The dimensionality of the model for one source is: $D + 3D(D+1)/2$ (mean vector and three covariance matrices), so $\mathfrak{D}(\boldsymbol{S}_{1:T,1:N}) = ND(3D+5)/2$.

2.1.2 Auditory status

The auditory states are represented in $\boldsymbol{A}_t$, which is a N-dimensional binary variable, i.e., $\boldsymbol{A}_t \in \{0,1\}^N, \forall t$. $A_{t,n} = 1$ if the n^{th} source is speaking at time t and $A_{t,n} = 0$ otherwise. Often, we abuse the notation by writing $\boldsymbol{A}_t = i$, being $i = 1, \dots, 2^N$ instead of the binary representation of $i - 1$. The auditory states are modelled as an HMM:

$$p\left(A_t = i | A_{t-1} = j, \mathbf{T}^A\right) = \mathbf{T}^A[i,j] \tag{2.3}$$

$$p\left(A_1 = i | \boldsymbol{\lambda}^A\right) = \boldsymbol{\lambda}^A[i]\,. \qquad i,j = 1, \dots, 2^N, \tag{2.4}$$

With $\theta^A = \left\{\mathbf{T}^A, \boldsymbol{\lambda}^A\right\}$, being $\mathbf{T}^A \in \mathcal{SM}^{2^N \times 2^N}$, the transition matrix and $\boldsymbol{\lambda}^A \in \mathcal{SV}^{2^N}$ the initial probabilities. The dimensionality of the model is $\mathfrak{D}(\boldsymbol{A}_{1:T}) = 2^N - 1 + 2^N(2^N - 1)$, from the stochastic vector and matrix, that is $\mathfrak{D}(\boldsymbol{A}_{1:T}) = 4^N - 1$.

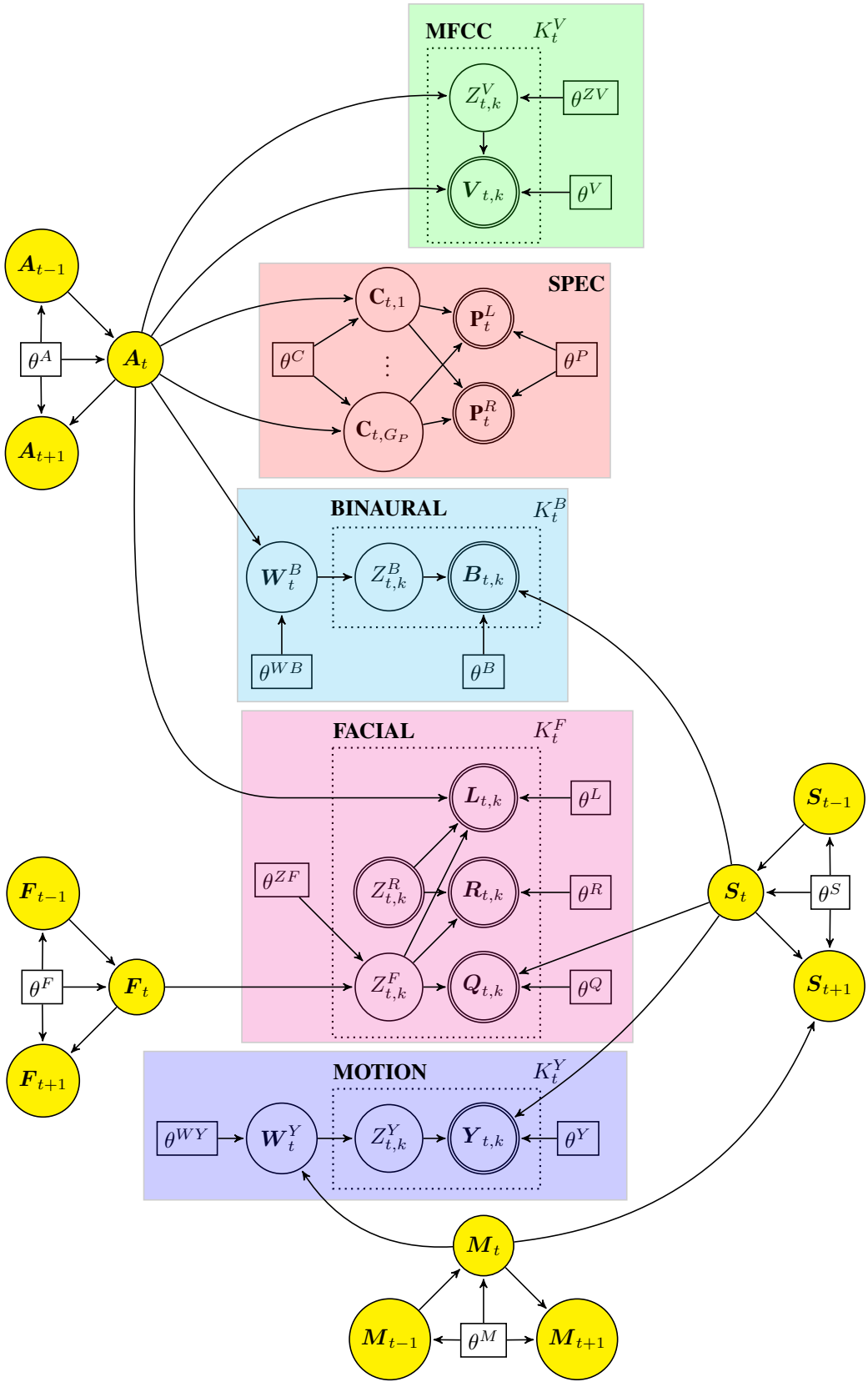

Figure 1: Full graphical model at time t. Rectangular nodes mean parameters, single circular nodes mean hidden variables and double circular nodes mean observations. Variables are grouped into six colors. The yellow ones, $\boldsymbol{A}_t$, $\boldsymbol{F}_t$, $\boldsymbol{M}_t$ and $\boldsymbol{S}_t$, correspond to dynamic variables, i.e., those which their values across time are linked. The green variables, $Z_{t,k}^V$ and $\boldsymbol{V}_{t,k}$, are related to the modelling of the Mel Frequency Cepstral Coefficients (**MFCC**). Red variables, $\mathbf{C}_{t,1}, \ldots, \mathbf{C}_{t,J}$, $\mathbf{P}_t^L$ and $\mathbf{P}_t^R$, model the spectrogram (**SPEC**). Cyan variables, $\boldsymbol{W}_t^B$, $Z_{t,k}^B$ and $\boldsymbol{B}_{t,k}$, model the extraction of sound source locates from binaural cues. (**BINAURAL**). Magenta variables, $Z_{t,k}^F$, $\boldsymbol{L}_{t,k}$, $\boldsymbol{R}_{t,k}$ and $\boldsymbol{Q}_{t,k}$ are related to the facial detections (**FACIAL**). Finally, blue variables, $\boldsymbol{W}_t^Y$, $Z_{t,k}^Y$ and $\boldsymbol{Y}_{t,k}$, model the motion features (**MOTION**).

2.1.3 Visibility

The visibility variable indicates whether the face of the n^{th} speaker is visible at time t ($F_{t,n} = 1$) or not ($F_{t,n} = 0$). They are modelled as independent HMM:

$$p\left(F_{t,n} = i | F_{t-1,n} = j, \mathbf{T}_n^F\right) = \mathbf{T}_n^F[i,j] \tag{2.5}$$

$$p\left(F_{1,n} = i | \boldsymbol{\lambda}_n^F\right) = \boldsymbol{\lambda}_n^F[i]\,. \tag{2.6}$$

We notice that $\theta^F = \left\{\mathbf{T}_1^F, \ldots, \mathbf{T}_N^F, \boldsymbol{\lambda}_1^F, \ldots, \boldsymbol{\lambda}_N^F\right\}$ with $\mathbf{T}_n^F \in \mathcal{SM}^{2\times 2}$ and $\boldsymbol{\lambda}_n^A \in \mathcal{SV}^2$, $\forall n = 1, \ldots, N$. Therefore, the dimensionality of the model is $\mathfrak{D}(\boldsymbol{F}_{1:T}) = N(2(2-1)) + N(2-1) = 3N$.

2.1.4 Motion

The motion variable indicates whether the n^{th} speaker is moving or not at time t, $M_{t,n} \in \{0, 1\}$.

$$p\left(M_{t,n} = i | M_{t-1,n} = j, \mathbf{T}_n^M\right) = \mathbf{T}_n^M[i,j] \tag{2.7}$$

$$p\left(M_{1,n} = i | \boldsymbol{\lambda}_n^M\right) = \boldsymbol{\lambda}_n^M[i]\,. \tag{2.8}$$

Therefore $\theta^M = \left\{\mathbf{T}_1^M, \ldots, \mathbf{T}_N^M, \boldsymbol{\lambda}_1^M, \ldots, \boldsymbol{\lambda}_N^M\right\}$, the same as in the previous case. Obviously, the dimensionality of the model is also $\mathfrak{D}(\boldsymbol{M}_{1:T}) = 3N$.

2.2 Modalities

2.2.1 Mel Frequency Cepstral Coefficients

The Mel Frequency Cepstral Coefficients (MFCC) are denoted by $\boldsymbol{V} \in \mathbb{R}^{D_V}$. The correspond to the green box in Figure 1, also shown in Figure 2. V are modelled using Gaussian Mixture Models (GMM). The set of parameters of the GMM, that is mixing proportions, mean vectors and covariance matrices, depend on the auditory status $\boldsymbol{A}_t$. We assume that K_t^V MFCC vectors are extracted at time t, and that they are independent and identically distributed realizations. GMMs have G_V components. For the auditory state $\boldsymbol{A}_t = j$, the GMM is parametrized by the mixing parameters $\boldsymbol{\pi}_j^V \in \mathcal{SV}^{G_V}$, the mean vectors $\boldsymbol{\mu}_{j,1}^V, \ldots, \boldsymbol{\mu}_{j,G_V}^V \in \mathbb{R}^{D_V}$ and the covariance matrices $\boldsymbol{\Sigma}_{j,1}^V, \ldots, \boldsymbol{\Sigma}_{j,G_V}^V \in \mathcal{CM}^{D_V}$, $\forall j = 1, \ldots, 2^N$. The hidden variable $Z_{t,k}^V$ represents the assignment of the k^{th} MFCC observation at time t, $\boldsymbol{V}_{t,k}$, to one of the mixture's components. Therefore, $Z_{t,k}^V \in \{1, \ldots, G_V\}$.

$$Z_{t,k}^V | \boldsymbol{A}_t = j \sim \text{Mult}\left(\boldsymbol{\pi}_j^V\right) \tag{2.9}$$

$$\boldsymbol{V}_{t,k} | \boldsymbol{A}_t = j, Z_{t,k}^V = g \sim \mathcal{N}\left(\boldsymbol{\mu}_{j,g}^V, \boldsymbol{\Sigma}_{j,g}^V\right) \tag{2.10}$$

We now denote $\theta^{ZV} = \boldsymbol{\pi}_{1:2^N}^V$ and $\theta^V = \left\{\boldsymbol{\mu}_{j,g}^V, \boldsymbol{\Sigma}_{j,g}^V\right\}_{j=1,g=1}^{2^N, G_V}$. The dimensionality of the MFCC model per each value of the auditory state variable $\boldsymbol{A}_t$ is $G_V(D_V + D_V(D_V+1)/2) = G_V D_V(D_V+3)/2$, so the dimensionality is: $\mathfrak{D}(\boldsymbol{V}_{1:T,1:K^V}) = 2^N G_V D_V(D_V+3)/2$.

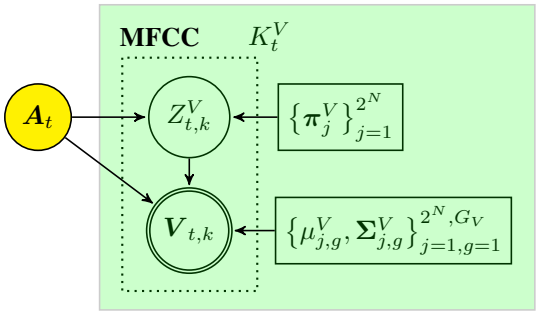

Figure 2: Graphical model of the MFCC. At each time interval we assume K_t^V i.i.d. realizations of the same GMM. The parameters of this GMM depend on the auditory state $\boldsymbol{A}_t$. The assignment variable $Z_{t,k}^V$ describes which of the G_V Gaussian components has generated the k^{th} observation $\boldsymbol{V}_{t,k}$. The parameters corresponding to this part of the model are the mixing coefficients $\boldsymbol{\pi}_j^V$, the mean vectors $\boldsymbol{\mu}_{j,g}^V$ and the covariance matrices $\boldsymbol{\Sigma}_{j,g}^V$.

2.2.2 Spectrogram

The spectrogram is modelled using mixtures of proper[1] complex Gaussians (see Figure 3). We inspired from the works of [3, 5]. The mixture has G_P components, assigned to the different speakers. Let $\{\mathcal{G}_n\}_{n=1}^N$ be a non-trivial partition of $\{1, \ldots, G_P\}$, then component g is assigned to source n if and only if $g \in \mathcal{G}_n$. Each Time-Frequency point of each component's spectrogram follows a proper complex Gaussian:

$$\mathbf{C}_{t,g}[f,i]\,|A_{t,n_g} \sim A_{t,n_g}\mathcal{N}_c\left(0, \boldsymbol{\omega}_g[f]\,\boldsymbol{\xi}_g[i]\right) \tag{2.11}$$

with $f = 1, \ldots, P_F$, $i = 1, \ldots, P_I$. n_g is the source to which the g^{th} component is assigned (i.e. $n_g = n \Leftrightarrow g \in \mathcal{G}_n$). The two vectors $\boldsymbol{\omega}_g \in (\mathbb{R}^+)^{P_F}$ and $\boldsymbol{\xi}_g \in (\mathbb{R}^+)^{P_I}$ model the power of the spectrogram of $\mathbf{C}_{t,g}$. The final left spectrogram is:

$$\mathbf{P}_t^L[f,i]\,|\mathbf{C}_{t,1:G_P}[f,i] \sim \mathcal{N}_c\left(\sum_{g=1}^{G_P} \boldsymbol{\Xi}^L[f,g]\,\mathbf{C}_{t,g}[f,i]\,, \boldsymbol{\Sigma}_f^{PL}\right), \tag{2.12}$$

where $\boldsymbol{\Xi}^L \in \mathbb{C}^{P_F \times G_P}$ is the mixing matrix and $\boldsymbol{\Sigma}_f^{PL} \in \mathcal{CM}^1 = \mathbb{R}^+$ is the noise variance. The same formula holds for the right spectrogram with $\boldsymbol{\Xi}^R$ and $\boldsymbol{\Sigma}_f^{PR}$. Notice that $\boldsymbol{\Xi}^L$ and $\boldsymbol{\Xi}^R$ are structured matrices. Indeed, the rows corresponding to the same source must be identical. Therefore, the parameters are: $\theta^C = \{\boldsymbol{\omega}_g, \boldsymbol{\xi}_g\}_{g=1}^{G_P}$ and $\theta^P = \{\boldsymbol{\Xi}^L, \boldsymbol{\Xi}^R, \{\boldsymbol{\Sigma}_f^{PL}, \boldsymbol{\Sigma}_f^{PR}\}_{f=1}^F\}$; and the dimensionality of the model is $\mathfrak{D}(\mathbf{P}_{1:T}^L, \mathbf{P}_{1:T}^R) = G_P(P_F + P_I) + 2(2G_P P_F + P_F)$, from the $\boldsymbol{\omega}$'s and the $\boldsymbol{\xi}$'s, the two mixing matrices $\boldsymbol{\Xi}^L, \boldsymbol{\Xi}^R$ and the noise $\boldsymbol{\Sigma}_f^{PL}, \boldsymbol{\Sigma}_f^{PR}$.

2.2.3 Binaural SSL cues

We will use binaural cues to help in the source localisation task. Binaural cues, denoted by $\boldsymbol{B}$ follow a GMM with a Dirichlet prior

[1] Proper complex Gaussians, also called circular symmetric complex Gaussians, have zero mean and zero relation matrix [4].

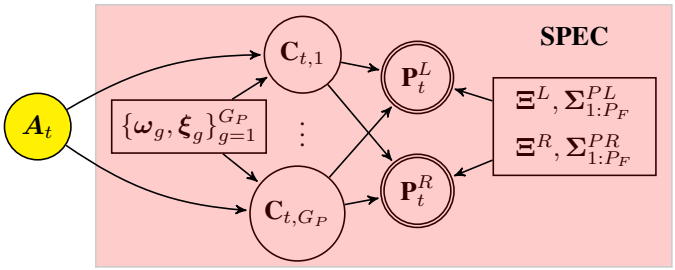

Figure 3: Graphical model of the spectrogram. The G_P components $\mathbf{C}_{t,g}$ are assigned to the different sources and the power spectrum at each time frequency point is modelled using $\boldsymbol{\omega}_g$ and $\boldsymbol{\xi}_g$. This hidden components are mixed with the matrices $\boldsymbol{\Xi}^L$ and $\boldsymbol{\Xi}^R$. Finally stationary noise is added in form complex Gaussian noise (with covariance matrices $\boldsymbol{\Sigma}_f^{PL}$ and $\boldsymbol{\Sigma}_f^{PR}$ respectively.

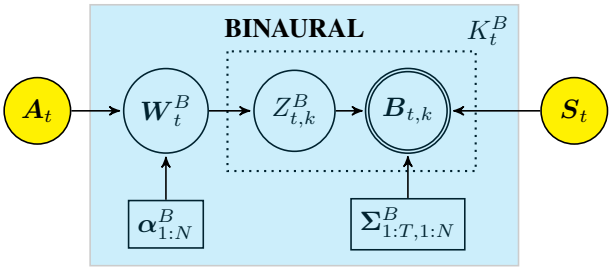

Figure 4: Graphical model for the binaural cues. $\boldsymbol{W}_t^B$ follows a Dirichlet distribution whose parameters, $\boldsymbol{\alpha}_{1:N}^B$, depend on the auditory state $\boldsymbol{A}_t$. One realization of the variable corresponds to the mixing coefficients for the GMM at time t. The means of the GMM correspond to the source positions $\boldsymbol{S}_{t,1:N}$ and the covariances, $\boldsymbol{\Sigma}_{1:T,1:N}^B$ need to be estimated.

on the mixing coefficients (see Figure 4). We assume K_t^B observations at time t, all i.i.d., following the same GMM. The variable $\boldsymbol{W}_t^B \in \mathcal{SV}^N$ follows a Dirichlet distribution whose parameters depend on the auditory state:

$$\boldsymbol{W}_t^B | \boldsymbol{A}_t \sim \mathcal{D}(\boldsymbol{\alpha}^B(\boldsymbol{A}_t)) \tag{2.13}$$

with

$$\boldsymbol{\alpha}^B(\boldsymbol{A}_t)[n] = \begin{cases} \boldsymbol{\alpha}_n^B[0] & \text{if } A_{t,n} = 0 \\ \boldsymbol{\alpha}_n^B[1] & \text{if } A_{t,n} = 1 \end{cases} \quad n = 1, \dots, N. \tag{2.14}$$

The vector $\boldsymbol{W}_t^B$ is used as mixing coefficients of the GMM [1] (Chapter 10), so the parameters of the multinomial distribution generating the assignment variable $Z_{t,k}^B \in \{1, \dots, N\}$:

$$Z_{t,k}^B | \boldsymbol{W}_t^B, \sim \text{Mult}(\boldsymbol{W}_t^B). \tag{2.15}$$

Finally, the observations $\boldsymbol{B}_{t,k} \in \mathbb{R}^D$ follow a Gaussian distribution:

$$\boldsymbol{B}_{t,k} | Z_{t,k}^B = n, \boldsymbol{S}_t \sim \mathcal{N}(\boldsymbol{S}_{t,n}, \boldsymbol{\Sigma}_{t,n}^B) \tag{2.16}$$

Therefore, $\theta^{WB} = \boldsymbol{\alpha}_{1:N}^B$ and $\theta^B = \boldsymbol{\Sigma}_{1:T,1:N}^B$; and the dimensionality of the model is $\mathfrak{D}(\boldsymbol{B}_{1:T,1:K^B}) = 2N + TND(D+1)/2$, from the parameters of the Dirichlet, plus the covariance matrices ($D(D+1)/2$ per person per time interval). If there are not enough samples to correctly estimate the covariance matrices, we will use one covariance per source (so the same for all time intervals).

2.2.4 Facial cues

Facial cues consist on four observations: face position $\boldsymbol{Q}$, face descriptor $\boldsymbol{R}$, pose detection Z^R and lip movement detection $\boldsymbol{L}$. We assume the existence of K_t^F face detections at interval t. The assignment variable, $Z_{t,k}^F$ indicates to which source the k^{th} detected face is assigned, and follows a multinomial with $N+1$ elements (since we account for false detections):

$$Z_{t,k}^F | \boldsymbol{F}_t \sim \text{Mult}\left(\frac{(1-e)F_{t,1}}{N_t^F}, \dots, \frac{(1-e)F_{t,N}}{N_t^F}, e\right), \tag{2.17}$$

being $N_t^F = \sum_{n=1}^N F_{t,n}$ the number of visible people and e the false error rate of the face detector.

The face is detected using the method proposed in [7]. This method provides the position of the face $\boldsymbol{Q}_{t,k}$, an estimation of the face pose $Z_{t,k}^R \in \{1, \dots, G_R\}$ and a set of G_F landmarks (the number may depend on the pose). The position of the face follows a Gaussian distribution:

$$\boldsymbol{Q}_{t,k} | Z_{t,k}^F = n, \boldsymbol{S}_t \sim \mathcal{N}\left(\boldsymbol{S}_{t,n}, \boldsymbol{\Sigma}^Q\right) \tag{2.18}$$

where $\boldsymbol{\Sigma}^Q$ describes the distribution of the localisation error of the face detector. Therefore $\theta^Q = \boldsymbol{\Sigma}^Q$ and the dimensionality of the model is $\mathfrak{D}_{\mathfrak{1}:\mathfrak{T},\mathfrak{1}:\mathfrak{R}^\mathfrak{d}}(\boldsymbol{Q}) = D(D+1)/2$.

The extracted landmarks are used to build a face descriptor which consists on the concatenation of the landmarks' positions together with a D_A-dimensional local descriptor per landmark. Therefore, $\boldsymbol{R}_{t,k} \in \mathbb{R}^{G_F(D+D_A)}$. This observation is modelled as a Gaussian, in which the geometric part (the landmark's positions) is independent of the appearance part and the appearance observations are independent between them. We denote the geometric part of $\boldsymbol{R}$ as $\boldsymbol{R}^G \in \mathbb{R}^{G_F D}$ and the appearance parts as $\boldsymbol{R}_g^A \in \mathbb{R}^{D_A}$, for $g = 1, \dots, G_F$. Therefore $\boldsymbol{R} = \left(\left(\boldsymbol{R}^G\right)^\top, \left(\boldsymbol{R}_1^A\right)^\top, \dots, \left(\boldsymbol{R}_{G_F}^A\right)^\top\right)^\top$, modelled by:

$$\boldsymbol{R}_{t,k}^G | Z_{t,k}^F = n, Z_{t,k}^R = m \sim \mathcal{N}\left(\boldsymbol{\mu}_{n,m}^G, \boldsymbol{\Sigma}_{n,m}^G\right) \tag{2.19}$$

$$\boldsymbol{R}_{t,k,g}^A | Z_{t,k}^F = n, Z_{t,k}^R = m \sim \mathcal{N}\left(\boldsymbol{\mu}_{n,m,g}^A, \boldsymbol{\Sigma}_{n,m,g}^A\right), \tag{2.20}$$

for $g = 1, \dots, G_F$. We inspired from [2], but for the time being we remove the priors on the matrices and means. Therefore $\theta^R = \left\{\boldsymbol{\mu}_{n,m}^G, \boldsymbol{\Sigma}_{n,m}^G, \{\boldsymbol{\mu}_{n,m,g}^A, \boldsymbol{\Sigma}_{n,m,g}^A\}_{g=1}^{G_F}\right\}_{n=1,m=1}^{N,G_R}$. In all, $\mathfrak{D}(\mathbf{R}_{1:T,1:K^F}) = NG_R(G_F D(G_F D+3) + G_F D_A(D_A+3))/2$.

The lip movement descriptor will consist on the concatenation of spatio-temporal features (e.g. Laptev or [6]). Therefore, we consider G_L (one per lip landmark) independent D_L-dimensional features all of them following a different Gaussian:

$$\boldsymbol{L}_{t,k,g} | Z_{t,k}^F = n, Z_{t,k}^R = m, A_{t,n} = a \sim \mathcal{N}\left(\boldsymbol{\mu}_{n,m,g,a}^L, \boldsymbol{\Sigma}_{n,m,g,a}^L\right), \tag{2.21}$$

with $g = 1, \dots, G_L$. Therefore $\theta^L = \left\{\boldsymbol{\mu}_{n,m,g,a}^L, \boldsymbol{\Sigma}_{n,m,g,a}^L\right\}_{n=1,m=1,g=1,a=0}^{N,G_F,G_L,1}$ and the dimensionality of $\boldsymbol{L}$: $\mathfrak{D}(\boldsymbol{L}_{1:T,1:K^F}) = 2NG_R G_L D_L(D_L+3)/2$.

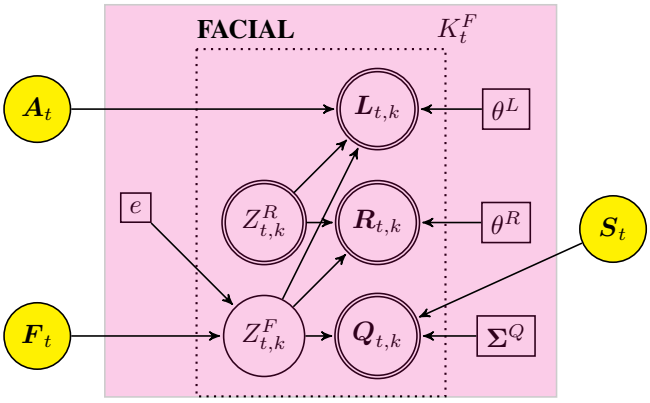

Figure 5: Graphical model of the facial cues. K_t^F faces are detected at time t. $Z^F_{t,k}$ assigns the faces to the sources, and $Z^R_{t,k}$ is the estimated pose. $\boldsymbol{Q}_{t,k}$ is the position of the face, $\boldsymbol{R}_{t,k}$ describes the geometry and the appearance of the face landmarks and $\boldsymbol{L}_{t,k}$ describes the local dynamics of the lip landmarks.

2.2.5 Movement cues

Movement cues have a model very similar to the binaural cues. We will use movement cues to help in the source localisation task. Motion cues, $\boldsymbol{Y}$ will be assumed to follow a GMM with a Dirichlet prior on the mixing coefficients (see Figure 6). We assume K_t^Y observations at time t, all i.i.d., following the same GMM. The variable $\boldsymbol{W}_t^Y \in \mathcal{SV}^N$ follows a Dirichlet distribution whose parameters depend on the auditory state:

$$\boldsymbol{W}_t^Y | \boldsymbol{M}_t \sim \mathcal{D}(\boldsymbol{\alpha}^Y(\boldsymbol{M}_t)) \tag{2.22}$$

with

$$\boldsymbol{\alpha}^Y(\boldsymbol{M}_t)[n] = \begin{cases} \boldsymbol{\alpha}_n^Y[0] & \text{if } M_{t,n} = 0 \\ \boldsymbol{\alpha}_n^Y[1] & \text{if } M_{t,n} = 1 \end{cases} \qquad n = 1, \ldots, N. \tag{2.23}$$

The vector $\boldsymbol{W}_t^Y$ is used as mixing coefficients of the GMM, so the parameters of the multinomial distribution generating the assignment variable $Z^Y_{t,k} \in \{1, \ldots, N\}$:

$$Z^Y_{t,k} | \boldsymbol{W}_t^Y, \sim \text{Mult}(\boldsymbol{W}_t^Y). \tag{2.24}$$

Finally, the observations $\boldsymbol{Y}_{t,k} \in \mathbb{R}^D$ follow a Gaussian distribution:

$$\boldsymbol{Y}_{t,k} | Z^Y_{t,k} = n, \boldsymbol{S}_t \sim \mathcal{N}(\boldsymbol{S}_{t,n}, \boldsymbol{\Sigma}^Y_{t,n}) \tag{2.25}$$

Therefore, $\theta^{WY} = \boldsymbol{\alpha}^Y_{1:N}$ and $\theta^Y = \boldsymbol{\Sigma}^Y_{1:T,1:N}$; and the dimensionality of the model is $\mathfrak{D}(\boldsymbol{Y}_{1:T,1:K^Y}) = 2N + TND(D+1)/2$, from the parameters of the Dirichlet, plus the covariance matrices ($D(D+1)/2$ per person per time interval). If there are not enough samples to correctly estimate the covariance matrices, we will use one covariance per source (so the same for all time intervals).

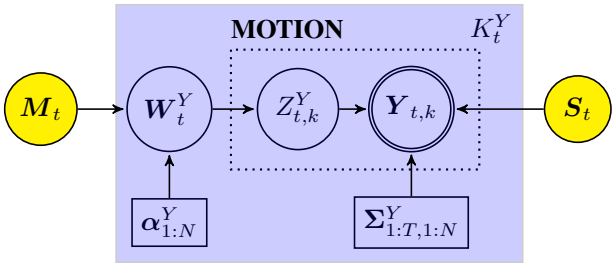

Figure 6: Graphical model for the binaural cues. $\boldsymbol{W}_t^Y$ follows a Dirichlet distribution whose parameters, $\boldsymbol{\alpha}^Y_{1:N}$, depend on the auditory state $\boldsymbol{M}_t$. One realization of the variable corresponds to the mixing coefficients for the GMM at time t. The means of the GMM correspond to the source positions $\boldsymbol{S}_t$ and the covariances $\boldsymbol{\Sigma}^Y_{1:T,1:N}$ need to be estimated.

References

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[2] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[3] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[4] NR Goodman. Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, 34(1):152–177, 1963.

[5] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):550–563, 2010.

[6] Guoying Zhao, Mark Barnard, and Matti Pietikainen. Lipreading with local spatiotemporal descriptors. *Multimedia, IEEE Transactions on*, 11(7):1254–1265, 2009.

[7] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.