Finite Mixture Spectrogram Modeling for Multipitch Tracking Using A Factorial Hidden Markov Model

Michael Wohlmayr, Franz Pernkopf

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
michael.wohlmayr@tugraz.at, pernkopf@tugraz.at

Abstract

In this paper, we present a simple and efficient feature modeling approach for tracking the pitch of two speakers speaking simultaneously. We model the spectrogram features using Gaussian Mixture Models (GMMs) in combination with the Minimum Description Length (MDL) model selection criterion. This enables to automatically determine the number of Gaussian components depending on the available data for a specific pitch pair. A factorial hidden Markov model (FHMM) is applied for tracking. We compare our approach to two methods based on correlogram features [1]. Those methods either use a HMM [1] or a FHMM [7] for tracking. Experimental results on the Mocha-TIMIT database [2] show that our proposed approach significantly outperforms the correlogram-based methods for speech utterances mixed at 0dB. The superior performance even holds when adding white Gaussian noise to the mixed speech utterances.

1. Introduction

Estimation and tracking of pitch is important for many algorithms and applications in speech and audio signal processing, e.g. one-channel blind source separation [3]. While well performing algorithms do exist for the case of a single speaker in a clean recording [4], the same task of pitch estimation is more difficult for noisy speech and multiple speakers talking simultaneously.

In [1], an approach for robust multipitch tracking has been proposed. It is based on the unitary model of pitch perception [5], upon which several improvements are introduced to yield a probabilistic representation of the periodicities in the signal. Semi-continuous pitch trajectories are then obtained by tracking these likelihoods using a hidden Markov model (HMM). Although this model provides an excellent performance in terms of accuracy, it is not possible to correctly link each pitch estimate to its source speaker.

Recently [8] and [7] show that factorial hidden Markov models (FHMMs) [6] provide a natural framework to track the pitch of multiple speakers. FHMMs enable to track the states of multiple Markov processes evolving in parallel over time, where the available observations are considered as a joint effect of all single Markov processes. Given these observations only, the task of inferring the most likely state sequence of each hidden Markov chain is more complex than for the HMM case. Although the junction tree algorithm [11] provides exact solutions to this problem, its computational complexity increases exponentially with the number of hidden Markov chains. Various methods to approximate the solutions for the sake of reduced complexity exist and can often be formulated as message passing algorithms [12].

2. Tracking with FHMM

A factorial HMM is a graphical model that employs several Markov chains in parallel. For simplicity, we present the case of two Markov chains, which is shown in Figure 1. The hidden state random variables are denoted by $x_{k}^{(t)}$, where $k$ indicates the Markov chain and $t$ the time index from 1 to $T$. Similarly, the observed random variables are denoted by $y^{(t)}$. Each $x_{k}^{(t)}$ represents a discrete random variable, while $y^{(t)}$ can be either discrete or continuous. For simplicity, all hidden variables are assumed to have cardinality $|X|$. The arcs between nodes indicate a direct conditional dependency between two random variables. Specifically, the dependency of hidden variables between two consecutive time instances is defined for each Markov chain by the transition probability $p(x_{k}^{(t)}|x_{k}^{(t-1)})$. The dependency of the observed variables $y^{(t)}$ on the hidden variables of the same time frame are defined by the obser-
Figure 1: A factorial HMM shown as a factor graph [13]. Factor nodes are shown as shaded rectangles together with their functional description. Hidden variable nodes are shown as circles. Observed variables $y^{(i)}$ are absorbed into factor nodes.

The term $p(y^{(i)}|x^{(i)}_1, x^{(i)}_2)$ is set to a small $\sigma_{min}$, where $I$ is the identity matrix. For $N > 1$, we train GMMs with $M$ ranging from 1 to 15, and take the GMM whose corresponding MDL criterion is minimal. If there is no training sample available, i.e., $N = 0$, $\mu_{m}^{x_{1}, x_{2}} = \bar{\mu}$ and $\Sigma_{m}^{x_{1}, x_{2}} = I$. Previously to pitch tracking all spectrogram samples are normalized to zero mean and unit variance. Finally, we multiply the pitch likelihood $p(y^{(i)}|x^{(i)}_1, x^{(i)}_2)$ with the pitch prior $p(x^{(i)}_1)p(x^{(i)}_2)$, since this slightly improved the performance in our experiments.

Both transition matrices of the FHMM $\Omega^{(k)} = p(x^{(k)}|x^{(k-1)})$ are obtained from training data. After initializing each entry to one, element $\Omega^{(k)}_{ij}$ is increased by one for each occurrence of a transition from state $j$ to state $i$ at speaker $k$. Finally, both matrices are normalized such that each column sums to one ($\sum_{j} \Omega^{(k)}_{ij} = 1$). Both prior distributions $p(x^{(i)}_1)$ are obtained likewise.

4. Tracking

The task of tracking involves searching the sequence of hidden states $x^*$ that maximizes the conditional distribution $p(x|y)$:

$$x^* = \arg \max_x p(x|y)$$

For HMMs, the exact solution to this problem is found by the Viterbi algorithm. For FHMMs, an exact solution can be found using the Junction tree algorithm [11], however this approach gets intractable for increasing $K$ and $|X|$. Several algorithms for approximate solutions are derived in [6] from the framework of variational inference. The Sum-Product algorithm [13] can be derived under a similar setting of variational principles [12], although more intuitive derivations exist for graphs without loops. When applied on a graph with loops, as is the case for a FHMM, the solutions are in general not guaranteed to converge and can only approximate the optimal solution.

In this work, we explore the Max-Sum (MS) algorithm (a variant of Sum-Product algorithm) to obtain a solution for equation 1. In [7], experimental results suggested that the obtained solutions sufficiently approximate the exact solution, while computational complexity is much lower. We give a short overview on the message passing algorithm used. For a detailed discussion, we refer the interested reader to [13, 11, 12]. The Max-Sum algorithm is based on passing messages between nodes of a graph. Among various types of graphs, factor graphs [13] have become a popular tool to depict the mechanisms of message passing. Figure 1 shows a FHMM as factor graph, where the functional dependency of
For the Max-Sum algorithm, each node sends to every neighbour a vector valued message $\mu$, which is itself a function of the messages it received, (as well as $f(x)$, for the case of a factor node). A message from variable node $x$ to factor node $f$ is

$$\mu_{x \rightarrow f}(x) = \sum_{y \in n(x) \setminus \{f\}} \mu_{y \rightarrow f}(x),$$ (2)

while a message from factor $f$ to variable $x$ is

$$\mu_{f \rightarrow x}(x) = \max_{y \in n(x) \setminus \{x\}} \left( \ln f(x) + \sum_{y \in n(f)} \mu_{y \rightarrow f}(y) \right) \text{ (3)}$$

Here, $n(x)$ denotes the set of neighbour nodes of $x$. We normalize each message and restrict each node to send a maximum of 15 messages per link. Further, each node only re-sends a message to a neighbour if it is significantly different from the previously sent message in terms of the Kullback-Leibler-divergence. After the last iteration, we obtain the maximum a posteriori configuration $p^*(x)$ of each variable node $x$ as a function of its incoming messages:

$$p^*(x) = \max_{x \in X} \mathcal{P}(x|y) = \sum_{y \in n(x)} \mu_{y \rightarrow x}(x) \text{ (4)}$$

Although the set of maxima $x^* = \arg\max_{x} p^*(x) \forall x \in X$ does not necessarily yield the global maximum $x^*$, as multiple global maxima might be present, a backtracking stage may lead to inconsistencies due to the loops in the factor graph. For this reason, we simply set the global maximum $x^*$ to the set of individual maxima $x^*$.

5. Experimental Results

The performance of MDL-FHMM was evaluated and compared with two other methods. The first method, proposed in [1], employs a sophisticated signal decomposition frontend, followed by a single HMM for tracking. We will refer to this method as COR-HMM. The second method, denoted as COR-FHMM, employs the same frontend as COR-HMM, however replaces the HMM with a FHMM [7].

Experiments have been performed using the "Mocha-TIMIT" database [2]. It consists of 460 English utterances from both a male and a female speaker, sampled at 16kHz. In addition, laryngograph signals are available for all recordings, from which the pitch ground truth $f_0[t]$ was acquired using the ESPS "get-0f" method [4] together with manual removal of erroneous pitch estimates in nonaudible regions. A training and test set of 400 and 60 sentences, respectively, was assembled, where each training and test instance was obtained by mixing a random male and female utterance at 0dB.

Training of our MDL-FHMM approach involves estimation of the transition and prior distributions, as well as the GMM estimation procedure described in section 3. In a first experiment, MDL-FHMM was compared with COR-HMM and COR-FHMM on the test set. For every test instance, each method estimates two pitch trajectories, $f_0^{(1)}[t]$ and $f_0^{(2)}[t]$. For performance evaluation, each of the two estimated pitch trajectories needs to be assigned to its ground truth trajectory, $f_0^{(1)}[t]$ or $f_0^{(2)}[t]$. From the two possible assignments, $(f_0^{(1)} \rightarrow f_0^{(1)}, f_0^{(2)} \rightarrow f_0^{(2)})$ or $(f_0^{(1)} \rightarrow f_0^{(2)}, f_0^{(2)} \rightarrow f_0^{(1)})$, the one is chosen for which the overall quadratic error is smaller. To evaluate the resulting estimates, we use the error measure proposed in [1]: $E_{ij}$ denotes the percentage of time frames where $i$ pitch points are misclassified as $j$ pitch points, i.e., $E_{12}$ means the percentage of frames with 2 pitches estimated whereas only one pitch is present. The pitch frequency deviation is defined as

$$\Delta f^{(k)}[t] = \frac{|\tilde{f}^{(k)}[t] - f_0[t]|}{f_0[t]}, \text{ (5)}$$
where \( f_0[t] \) denotes the reference chosen for \( \hat{f}_{(k)}[t] \). The gross detection error rate \( E_{\text{Gross}} \) is the percentage of time frames where the algorithm correctly detects the presence of one pitch (two pitches), but the corresponding frequency deviation \( \Delta f_{(k)}[t] \) (either of \( \Delta f^{(1)}[t] \) or \( \Delta f^{(2)}[t] \)) is larger than 20%. The fine detection error \( E_{\text{Fine}} \) is the average frequency deviation in percent at time frames where \( \Delta f_{(k)}[t] \) is smaller than 20%. The overall error, \( E_{\text{Total}} \), is defined as the sum of all error terms:

\[
E_{\text{Total}} = E_{01} + E_{02} + E_{10} + E_{12} + E_{20} + E_{21} + E_{\text{Gross}} + E_{\text{Fine}}.
\]

Table 1 shows the error measure for all 3 methods on the test set. MDL-FHMM significantly outperforms the correlation based methods COR-FHMM and COR-HMM. In a second experiment, each example of the test set was mixed with white noise at different SNR conditions, ranging from 40 dB down to 0 dB steps. For each SNR condition, the performance of MDL-FHMM was compared with COR-FHMM and COR-HMM, where the parameters of all methods were optimized for clean speech. The resulting performance of all methods is shown in figure 3. This figure shows that both correlation based methods are less robust for multipitch tracking of noise corrupted speech mixtures. Figure 2 shows tracking results for one test example corrupted with white noise at 0 dB SNR.

### 6. Conclusions

We have investigated the performance of a simple and effective modelling approach for multipitch tracking. First, spectral features are modelled with a set of GMMs, where the optimal complexity of each GMM is found via the MDL criterion. The set of GMMs then provides a basis for tracking the pitch of each speaker with a FHMM. The proposed method is superior in comparison to two other systems based on correlogram features.

### 7. Acknowledgements

This work was funded by the Austrian Science Fund (project number P19737-N15).

### 8. References


