GAIN-ROBUST MULTI-PITCH TRACKING USING SPARSE NONNEGATIVE MATRIX FACTORIZATION

Robert Peharz, Michael Wohlmayr, Franz Pernkopf

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

ABSTRACT

While nonnegative matrix factorization (NMF) has successfully been applied for gain-robust multi-pitch detection, a method to track pitch values over time was not provided. We embed NMF-based pitch detection into a recently proposed pitch-tracking system, based on a factorial hidden Markov model (FHMM). The original system models speech spectra with Gaussian mixture models, which is sensitive to a gain mismatch between training and test data. We therefore combine the advantages of these two approaches and derive a gain-adaptive observation model for the FHMM. As training algorithm we use a modification of ℓ0-sparse NMF, which represents the short-time spectrum with scalable basis vectors. In experiments we show that the new approach significantly increases the gain-robustness of the original tracking system.

Index Terms— multi-pitch, factorial model, sparse NMF

1. INTRODUCTION

The fundamental frequency of speech is an important cue for automatic speech recognition and prosodic analysis. Furthermore, when several speakers are interfering in a recording, the extraction of their fundamental frequencies can be considered as a sub-problem for single channel speech separation and computational auditory scene analysis [1]. The term pitch denotes the frequency of a sinusoid which is perceived to have the same tone as the considered signal, which is not necessarily identical to the fundamental frequency. However, to be consistent with literature, we use the term pitch instead of fundamental frequency.

Two of the best performing single pitch extraction methods are RAPT [2] and YIN [3]. For multi-pitch tracking, the algorithm described in [4] can be considered to be state-of-the-art. Although this algorithm extracts the pitch trajectories with high accuracy, it lacks in robustly assigning them to individual speakers. The probabilistic approach described in [5] explicitly models speech spectra using Gaussian mixture models (GMM), and uses an interaction model together with a factorial hidden Markov model to extract the most likely pitch-trajectories. This method significantly reduces permutation errors when speaker dependent models are used. However, the GMMs are trained on speech at a specific gain level and perform poorly in the case of a gain-mismatch to the test data. Although a global mismatch between training and test data can be corrected by adjusting the energy of the speech mixture, a correction of an imbalance among speakers is a non-trivial problem. However, for this method no mechanism was provided to track pitch estimates over time, nor to assign pitch values to individual speakers.

In this paper, we embed the gain-robust NMF approach [6] for multi-pitch detection into the probabilistic tracking system [5]. For this purpose, we define a scalable interaction model and propose to use non-negative least squares for gain estimation during inference. Our system therefore models merely the shape of speech spectra and intentionally ignores gain information in the training data. We show that classical NMF can not be used as training algorithm, when more than just one basis vectors per pitch value are desired, due to the naturally part-based representation of NMF [7] (cf. section 3). However, we circumvent this problem by using a modification of ℓ0-sparse NMF [8], where we enforce the algorithm to represent each spectrum with a single basis vector. We experimentally compare the original multi-pitch tracking system using GMMs [5] to our gain-robust variant. The proposed method has a significant advantage over the original approach when the speaker’s gains differ from each other, and achieves almost the same performance in the gain-balanced case.

The paper is organized as follows: Section 2 reviews the probabilistic system for multi-pitch tracking. In section 3 we discuss NMF [7, 9] and its application to multi-pitch detection [6]. In section 4 we present our modified sparse NMF algorithm. In section 5 we introduce the gain-adaptive observation model. Section 6 presents experimental results and section 7 concludes the paper.

2. PROBABILISTIC MULTIPITCH TRACKING

In this section, we shortly review the approach described in [5] for the case of two interfering speakers. The pitch trajectory of each speaker is modeled by a Markov chain of discrete states. Specifically, \( x_i^{(t)} \) denotes the pitch state of speaker at time frame \( t \), where \( x_i^{(t)} \in \{1 \ldots 170\} \). State ‘1’ denotes ’no pitch’, i.e. unvoiced speech or silence, and states ‘2’-’170’ encode specific pitch values according to \( f_0(x) = \frac{14600}{x} \) Hz, \( x \in \{2 \ldots 170\} \). This non-uniform quantization results in a finer resolution for low pitch values [4]. The prior probabilities \( p(x_1^{(t)}) \) and transitions \( p(x_1^{(t+1)}|x_1^{(t)}) \) are obtained from maximum-likelihood estimates on pitch sequences extracted from a corpus. The pitch dependent short-time magnitude spectrum \( s_i^{(t)} \) for speaker \( i \) is modelled using a GMM:

\[
p(s_i^{(t)}|x_i^{(t)}; \Theta_i, x_i) = \sum_{k=1}^{K_{i,x_i}} \alpha_{i,x_i,k}^k \mathcal{N}\left(s_i^{(t)}; \mu_{i,x_i,k}^k, \Sigma_{i,x_i,k}^k \right),
\]

where \( K_{i,x_i} \) is the number of Gaussian components for speaker \( i \) in pitch state \( x_i, \Theta_i, x_i = \{\alpha_{i,x_i,k}^k, \mu_{i,x_i,k}^k, \Sigma_{i,x_i,k}^k\}_{k=1}^{K_{i,x_i}} \) are the cor-

This work was supported by the Austrian Science Fund (project number S10604-N13 and P22488-N23).
responding prior probability, mean, and (diagonal) covariance matrix of the $k^{th}$ component, respectively. The GMM parameters are learned from labeled data using the expectation-maximization algorithm [10], and the number of components is selected using the minimum description length criterion [11].

The individual Markov chains – one for each speaker – are combined into a factorial hidden Markov model (FHMM). In the FHMM framework, the Markov chains evolve independently and jointly create the observations (i.e. the speech mixture), according to an interaction model. Two interaction models have been proposed [5], namely the linear interaction model for magnitude spectra, and the MIXMAX interaction model for magnitude-log spectra. Here we only consider the linear interaction model, where the mixture magnitude spectrum $y(t)$ is assumed to be the sum of the source magnitude spectra: $y(t) \approx s_1(t) + s_2(t)$. Taking use of the fact that the probability density function (pdf) of a sum of independent random variables is the convolution of the individual pdfs [12], we obtain the observation probability as

$$p(y(t) | x_1, x_2; \Theta_1, \Theta_2, x_{1,2}) = \sum_{m=1}^{M_1} \sum_{n=1}^{M_2} \alpha_{1,m}^{n} \alpha_{2,m}^{n} \mathcal{N}(y(t) | \mu_{1,m}^{n} + \mu_{2,m}^{n}, \Sigma_{1,m}^{n} + \Sigma_{2,m}^{n}).$$

The pitch state priors $p(x_1(t))^0$ and the observation probability $p(y(t) | x_1, x_2)$ fully specify the FHMM. Using a Viterbi decoder for FHMMs [5], we are able to infer the most likely pitch trajectories, given the observed speech mixture.

A major shortcoming in this approach is that possible gain-mismatches between training and testing are not treated. Reasons for such mismatches can be a changing distance between speaker and microphone, varying microphone amplifications, naturally varying loudness of the speaker, etc. A remedy for this problem would be to introduce explicit gain-parameters in the individual GMMs, and to adapt these parameters in terms of maximum-likelihood. While maximum-likelihood adaption is more or less straightforward for a single GMM and a single speech signal, it is non-trivial for the FHMM approach and speech mixtures. The standard approach, using expectation-maximization, in principle requires full inference for each M-step. For the current system, this approach is computationally infeasible, and will be treated in future work.

3. NONNEGATIVE MATRIX FACTORIZATION

Nonnegative matrix factorization aims to decompose a nonnegative matrix $X$ into a product of nonnegative matrices $W$ and $H$: $X \approx WH$. Lee and Seung [9] provided the multiplicative update rules

$$W \leftarrow W \odot \frac{XH^T}{WH^TH^T}, \quad H \leftarrow H \odot \frac{X^TW}{W^TWH^T}$$

which reduce the Frobenius norm $\|X - WH\|^2_F$, where $\odot$ and $-$ denote element-wise product and element-wise division, respectively. In general, we can interpret the columns of matrix $W$ as basis vectors generating the columns of the data matrix $X$, while $H$ represents the corresponding weight coefficients.

Sha and Saul [6] applied NMF to single channel multi-pitch detection, by training NMF bases on an instantaneous frequency spectrogram of pitch-labeled speech data. To extract pitch estimates from a mixture, they inferred weight coefficients for the trained NMF bases using nonnegative deconvolution, which is done by executing the NMF update rule for $H$ alone, keeping $W$ fixed. The bases with significant coefficients are regarded as detected pitch values. Contrary to the approach presented in section 2, the NMF approach does not suffer from a potential gain-mismatch between training and test data. However, this method does not provide a mechanism to track pitch values over time, nor does it assign them to specific speakers. Furthermore, in the work described in [6] only a single NMF basis vector is trained for each pitch value, which is quite limiting and does not cover the variety of speech data. An appealing way to generalize NMF-based pitch detection is to train several NMF bases for each pitch value. However, as it was pointed out in [7], NMF naturally finds a sparse and part-based representation of the input data. Indeed, it is this property, amongst others, which contributed to the success of NMF during the last decade. Hence, when applied to speech spectra of the same fundamental frequency, NMF tends to detect individual spectral peaks (cf. fig. 1). This behavior, however, is not desirable for multi-pitch detection, since individual spectral peaks are not well-suited to discriminate between different pitch values. In order to avoid a part-based representation, we can enforce sparseness on the columns of $H$. We consider the case where we allow only a single entry in each column of $H$ to be non-zero, i.e. each column of $X$ is represented by the scaled version of a single column out of $W$. For this purpose we use an $\ell^0$-sparse version of NMF [8], which is introduced in the next section.

4. SPARSE NONNEGATIVE MATRIX FACTORIZATION

Given a non-negative $D \times N$ data matrix $X$, the problem of sparse NMF can be formulated as finding non-negative matrices $W$ and $H$ minimizing $\|X - WH\|^2_F$, under the constraints $L_0(h_n) \leq L, \forall n$, where $h_n$ is the $n^{th}$ column of $H$ and $L_0(\cdot)$ denotes the $\ell^0$-pseudo-norm (i.e. the number of non-zeros entries). This means that each column of $X$ is approximated by a non-negative linear combination of maximal $L$ basis vectors. The dimensions of the matrices $W$ and $H$ are given as $D \times K$ and $K \times N$, respectively, where $K$ is the inner approximation rank or number of basis vectors. An iterative procedure was proposed for sparse NMF [8], consisting of the following two stages:

1. minimize $\|X - WH\|^2_F$, s.t. $L_0(h_n) \leq L, \forall n$, with respect to $H$, keeping $W$ fixed.
2. execute several iterations of classical NMF (cf. (3)).

Step 2 enhances the basis matrix $W$ and adapts the coefficients, while maintaining the sparse structure of $H$, due to the sparseness-maintaining property of classical NMF [8]. Step 1 is NP-hard in
general and has to be solved approximately by a non-negative sparse coder. However, for the special case of $L = 1$, we can make several simplifications to the algorithm. First, step 1 can be solved exactly by non-negative matching pursuit [8], which is described in algorithm 1. Without loss of generality, we assume that the columns of $W$ are normalized to unit length. $h_i$ denote the entries of $H$. Second, since $HH^T$ is diagonal for $L = 1$, $W$ cancels in (3, left)

**Algorithm 1** Non-negative Matching Pursuit ($L = 1$)

1. for $n = 1 : N$ do
2. $\mathbf{a} = \mathbf{W}^T \mathbf{x}_n$
3. $i = \arg\max \mathbf{a}$
4. $h_{i,n} = \max \mathbf{a}$
5. $h_{i,n} = 0, \forall j \neq i$
6. end for

and we obtain the simplified NMF update rule $W \leftarrow \frac{XH^T}{HH^T}$, where $1$ is a $D \times K$ matrix containing only ones. This update is a one-step solution, since it does not depend on $W$ any more. Furthermore, since $HH^T$ is diagonal, the new update can be written as $W \leftarrow XH^T(HH^T)^{-1}$, which formally recognize as Moors Penrose inverse. The sparse NMF algorithm, for the special case $L = 1$, is summarized in algorithm (2), where $I$ is the number of iterations. Step 5 assures that the NMF bases are normalized to unit length, as required by algorithm 1. Note that this algorithm has great similarity to k-means. Indeed, we can interpret the columns of $W$ as scalable ‘cluster centers’, where scale and data-to-‘cluster’ assignment is encoded by the value and the position of the non-zero entries in $H$, respectively.

5. **GAIN ADAPTIVE OBSERVATION MODEL**

We now introduce a scalable model for speaker dependent short-time spectra, where for simplicity we omit the dependency on time frame $t$:

$$p(s_i|x_i; \Theta_{i,x}) = \sum_{k=1}^{K_{i,x}} \alpha_{k}(s_i) \mu_{i,x}^k g_{i,x}(s_i) \lambda_{i,x}^k.$$  (4)

Here $L$ denotes the multivariate Laplacian distribution defined as $L(x | \mu, \lambda) = \prod_{d=1}^D \frac{1}{\lambda_d} e^{-\frac{|x_d - \mu_d|}{\lambda_d}}$, where $x \in \mathbb{R}^D$, and $\mu$ and $\lambda$ are mean and width parameters, respectively. Similar as in (1), $s_i$ is the short-time magnitude spectrum of speaker $i$, $K_{i,x}$ denotes the number of components for speaker $i$ in pitch state $x_i$, and $\alpha_k(s_i)$ is the respective component prior. We introduced a gain parameter $g_{i,x}^k$ for each component, which scales the component’s mean $\mu_{i,x}^k$ and width $\lambda_{i,x}^k$. We use Laplacian distributions instead of Gaussians, since we observed that the residual error $E = X - WH$ of algorithm 2 is distributed similar to a zero-mean Laplacian. Furthermore, we consistently obtained a slightly better performance in our experiments using Laplacians instead of Gaussians. The parameters $\Theta_{i,x} = \{\alpha_k(s_i), \mu_{i,x}^k, \lambda_{i,x}^k \}^k_{k=1}$ are obtained by applying algorithm 2 to a matrix $X$, containing short-time spectra of speaker $i$ with pitch state $x_i$. The number $K_{i,x}$ of trained basis vectors is determined as $K_{i,x} = \min\left(\left\lceil \frac{N}{50} \right\rceil, 20\right)$, where $N$ is the number of training examples available for pitch state $x_i$ (i.e. the number of columns in $X$), and $[.]$ is the floor operator. The mean vectors $\mu_{i,x}^k$ are the resulting bases in $W$, and the component prior $\alpha_k(s_i)$ is the relative frequency of $(h_{i,n})$ being non-zero. For a Laplacian random variable with variance $v$, the width parameter is given as $\lambda = \sqrt{2v}$. To estimate $\lambda_{i,x}^k$, we first normalize the data according to $s_n = \frac{x_n}{\max \{x_n\}}, n \in \{1 \ldots N\}$, where $\max \{|x_n|\}$ is equal to the value of the unique non-zero entry in $h_n$. We split $X$ into $K$ matrices $X^k, k \in \{1 \ldots K\}$, where $X^k$ contains all columns of $X$ associated with $w_k$ (i.e. where $h_{i,n} > 0, n \in \{1 \ldots N\}$). Now we can estimate the $d$th entry of $\lambda_{i,x}^k$ using

$$\lambda_{i,x}^k(d) = \frac{1}{2(N - 1)} \sum_{n=1}^{N-1} \left(\frac{x_{d,n} - w_{d,k}}{\alpha_k(s_i)} \right)^2, \quad (5)$$

where $N_k$ is the number of columns in $X^k$. Using the linear interaction model, i.e. assuming $y \approx y_1 + y_2$, and proceeding similar as in section 2, we obtain the scalable observation model:

$$p(y|x_1, x_2, \Theta_{1,x_1}, \Theta_{2,x_2}) = \sum_{k=1}^{K_{i,x_1}} \sum_{l=1}^{K_{i,x_2}} \alpha_k(s_i) \alpha_{l}(s_i) \tilde{L}_2 \left( y | \mu_{1,x_1}^k, \mu_{2,x_2}^l, \lambda_{1,x_1}^k, \lambda_{2,x_2}^l \right). \quad (6)$$

With $\tilde{L}_2$ we denote the convolution of two Laplacians, given as (see [13] for a derivation):

$$\tilde{L}_2(y | \mu_1, \mu_2, \lambda_1, \lambda_2) = \prod_{d=1}^D \frac{1}{2} \left[ -\frac{(\lambda_1 d)^2 - (\lambda_2 d)^2}{(\lambda_1 d)^2 - (\lambda_2 d)^2} - \frac{y_d - \mu_1 - \mu_2}{\lambda_1 d} \right]. \quad (7)$$

While the parameters $\Theta_{i,x}$ are obtained from algorithm 2, it remains unclear how to choose the gain parameters $g_{i,x_1}$ and $g_{i,x_2}$. In a Bayesian approach, we would introduce a prior and marginalize over $g_{i,x_1}$ and $g_{i,x_2}$. However, all parameters in (7) are dependent on one of the gain parameters, such that we can not find an analytic solution for this problem. Instead, we find point-estimates for $\tilde{g}_{i,x_1}$ and $g_{i,x_2}$ for each time frame $t$, by fitting $\mu_{1,x_1}^k$ and $\mu_{2,x_2}^l$ to $y$ in a non-negative least squares sense:

$$\left( \tilde{g}_{i,x_1}^k, g_{i,x_2}^l \right) = \arg\min_{h_1, h_2} \left\| y - (h_1 \mu_{1,x_1}^k + h_2 \mu_{2,x_2}^l) \right\|_2^2, \quad \text{s.t. } h_1, h_2 \geq 0 \quad (8)$$

Note that this is equivalent to find NMF coefficients $H = (h_1, h_2)^T$ for the fixed matrix $W = (\mu_{1,x_1}^k, \mu_{2,x_2}^l)$. Using the gains determined by (8), we can evaluate the scalable observation likelihood $p(y|x_1, x_2; \Theta_{1,x_1}, \Theta_{2,x_2})$ defined in (6). Together with the state priors and transitions described in section 2, we determine the most likely pitch trajectories using the Viterbi decoder for FHMMs [5].
Fig. 2. Tracking result as a function of gain level, for the original system using GMMs, the gain-robust system using NMF, and the multipitch tracker by Wu et al. [4]. The lines correspond to the median of $E_{\text{Total}}$ over all mixture utterances. The shaded bars correspond to the 25% and 75% quantiles.

6. EXPERIMENTS

We compared the gain adaptive observation model to the original GMM model as described in [5], and to the multi-pitch tracker by Wu et al. [4]. We selected 4 speakers (2 female, 2 male) from the GRID data base [14]. For each speaker we selected 10 utterances for testing and 490 utterances for GMM and NMF training. The speech signals were sampled at $f_s = 16$ kHz and the spectrograms were calculated using a Hamming window of length 1024 with 50% overlap and a 1024 point FFT. The observations were obtained by taking the magnitude of frequency bins 2-65, which corresponds to a frequency range up to 1 kHz. The performance of the algorithms was measured using $E_{\text{Total}}$ as defined in [5]. Note that $E_{\text{Total}}$, unlike $E_{\text{Total}}$ [4], additionally measures permutation errors, in order to evaluate model-based approaches such as the two systems discussed in this paper. The pitch tracker described in [4] is not model-based. Hence, it is prone to permutation errors and should be understood as baseline for this experiment.

We considered all possible combinations of utterances, which gives 600 mixture configurations in total. For each mixture configuration, we alternately amplified one of the two speakers with a gain level out of $G = \{-18, -12, -6, 0, 6, 12, 18\}$ dB, which gives 1200 mixture utterances per gain level (600 for gain level 0 dB). Figure 2 shows the tracking result in terms of $E_{\text{Total}}$ for all methods as a function of the gain parameter. The system using GMMs performs slightly better than the NMF-based method for the 0 dB mixing level, since in this case both GMMs optimally match the speakers’ gain. In all other cases the gain-adaptive system performs significantly better. Furthermore, the tracking performance of the NMF-based system remains reasonable robust, even for a relative gain mismatch of $\pm 18$ dB. Note that the curves for the NMF-based approach and the method by Wu et al. are symmetric, since the set of mixtures for gain level $+G$ is identical as for $-G$, up to an overall mixture gain. Both methods, however, are not affected by an overall gain level. On the other hand, the GMM-based approach suffers more for positive gain levels than for negative ones. For positive gain levels, the mixture is dominated by the amplified speaker (whose GMM is not well-adapted), which also misleads the GMM of the unaltered speaker.

7. CONCLUSION

In this paper we incorporated NMF-based multi-pitch detection into a probabilistic framework for multi-pitch tracking based on a FHMM. The proposed system models the shape of speech spectra using gain-adaptive NMF bases. Hence it infers the pitch-trajectories using merely the spectral shape and intentionally ignores any gain information available in the mixture signal. As a direct consequence, the system is not affected by an overall mixture gain level. Moreover, our system performs almost as well as the GMM approach (which implicitly uses gain information), when the GMMs are optimal adapted, and performs significantly better in the unbalanced case. Moreover, the new approach shows a robust performance over a large range of different mixing levels among the speakers.

8. REFERENCES