MODEL ADAPTATION OF FACTORIAL HMMS FOR MULTIPITCH TRACKING

M. Wohlmayr and F. Pernkopf

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

ABSTRACT

Factorial hidden Markov models (FHMMs) are used for tracking the pitch of two interacting speakers [1]. In this statistical approach, the characteristics of each speaker are captured by pre-trained models. Speaker models that match the test conditions well allow for high tracking performance, however the availability of such models is unrealistic. To extend the applicabiliy of the FHMM framework, we develop an EM-like iterative adaptation algorithm which is capable to adapt the model parameters to the specific situation, e.g. acoustic channel, using only speech mixture data. Model adaptation is empirically evaluated using real room recordings of mixture utterances from the GRID corpus.

***Index Terms*—** Multipitch tracking, factorial HMMs, self-adaptation, MLLR

1. INTRODUCTION

Recently, a system based on FHMMs and speaker interaction models has won the monaural speech separation and recognition challenge [2, 3]. Remarkably, this model slightly outperforms human listeners on a restricted task [4]. Independently, we developed a similar FHMM model for multipitch tracking [1]. These FHMM models are well-suited for modeling acoustic scenes of multiple interacting sources. However, these models require speaker/source specific data for learning which limits the applicability. Some of the most successful approaches for model adaptation in the context of speech recognition are the maximum likelihood linear regression (MLLR) framework [5, 6], maximum a posteriori (MAP) estimation [7], and rapid adaptation in eigenvoice space [8]. While these approaches assume that adaptation data consists of clean speech, methods for adaptation of *undistorted* source models from contaminated speech have been also developed, e.g. in [9], speech and noise – separately represented by individual models – are combined using the MIX-MAX model. Rose et al. [10] extended this using a more general interaction and background noise model based on Gaussian mixture models (GMMs). In [11], the eigenvoice approach is generalized to adapt individual speaker models given a superposition of two speech signals. Other approaches are summarized in detail in [12].

In this paper, we develop an algorithm for model adaptation to overcome any mismatch between training and testing conditions. The aim is to adapt *universal* models learned on data from many speakers to a novel acoustic environment using only speech mixture data. We propose an EM-based iterative algorithm using MLLR for adaptation of speaker models from speech mixtures, and demonstrate multipitch tracking[1] results obtained for a distant talking scenario of two speakers which includes room reverberation. Our approach overcomes the necessity of clean source-specific data for model adaptation. Furthermore, we constrain MLLR to modifications on the spectral envelope only which is beneficial in cases of few adaptation data.

The paper is organized as follows: In Section 2 we introduce FHMMs and the speaker interaction model. Section 3 presents the EM-framework for model adaptation. In Section 4 empirical results are reported. Section 5 concludes with a perspective on future work.

2. FHMMS WITH INTERACTION MODELS

FHMMs are capable to model a mixture of several speakers as a joint effect of multiple Markov processes evolving in parallel over time.[2] By combining two single speaker HMMs using an interaction model, we obtain the FHMM shown in Figure 1.[3]

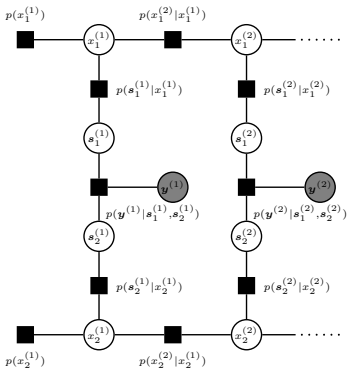

Fig. 1. FHMM represented as factor graph. The single-speaker emissions $\boldsymbol{s}_1^{(t)}$ and $\boldsymbol{s}_2^{(t)}$ jointly produce the observation $\boldsymbol{y}^{(t)}$.

The hidden random variables (RVs) $x_k^{(t)}$ represent the pitch state at time t of speaker k. The discrete hidden variable $x^{(t)}$ has $|X| = 170$ states, where state value '1' refers to 'no pitch' (i.e. unvoiced speech or silence), and state values '2'-'170' encode different pitch frequencies ranging from 80 to 500Hz. As in [1], the pitch value of state $x \in \{2, ..., 170\}$ is $f_0 = \frac{16000}{30+x}$. Vector $\boldsymbol{s}_k^{(t)} \in \mathbb{R}^D$ corresponds to D bins of the short-time log-magnitude DFT of speaker k at time frame t. The dependency of $x_k^{(t)}$ over time is modeled by

This work was supported by the Austrian Science Fund (FWF) under the project number P25244-N15 and S10610-N13.

[1]We track the fundamental frequency f_0. But we use the term pitch, usually used in psycho-acoustics, since it is more consistent with previous literature [13, 1, 14].

[2]For simplicity we consider the case of two interfering speakers throughout this section. This is generalized to K speakers subsequently.

[3]Factor nodes are depicted as shaded rectangles together with their functional description. Hidden variable nodes are shown as white circles, observed variable nodes as gray circles.

the transition probability $p(x_k^{(t)}|x_k^{(t-1)})$ and the prior distribution is denoted by $p(x_k^{(1)})$. The dependency of $\boldsymbol{s}_k^{(t)}$ on $x_k^{(t)}$ is modeled as GMM according to[4]

$$p(\boldsymbol{s}_k|x_k) = p(\boldsymbol{s}_k|\boldsymbol{\Theta}_{k,x_k}) = \sum_{m=1}^{M_{k,x_k}} \alpha_{k,x_k}^m \mathcal{N}\left(\boldsymbol{s}_k|\boldsymbol{\theta}_{k,x_k}^m\right), \quad (1)$$

where $M_{k,x_k} \geq 1$ is the number of mixture components of speaker k and state x_k, and α_{k,x_k}^m denotes the weight of component m; $\alpha_{k,x_k}^m \geq 0$ and $\sum_{m=1}^{M_{k,x_k}} \alpha_{k,x_k}^m = 1$. The GMM for state x_k is specified by the parameter set $\boldsymbol{\Theta}_{k,x_k} = \left\{\alpha_{k,x_k}^m, \boldsymbol{\theta}_{k,x_k}^m\right\}_{m=1}^{M_{k,x_k}}$, where $\boldsymbol{\theta}_{k,x_k}^m = \left\{\boldsymbol{\mu}_{k,x_k}^m, \boldsymbol{\Sigma}_{k,x_k}^m\right\}$ is the mean and diagonal covariance of component m. To keep the notation compact, we use braces to denote a set of RVs from all Markov chains, e.g. $\{x_k^{(t)}\} := \{x_k^{(t)}\}_{k=1}^K$. At each time frame, the observation $\boldsymbol{y}^{(t)}$ is considered to be produced jointly by the two single-speaker emissions $\boldsymbol{s}_1^{(t)}$ and $\boldsymbol{s}_2^{(t)}$ using the mixture-maximization (MIXMAX) model [9] $p(\boldsymbol{y}^{(t)}|\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)}) = \delta\left(\boldsymbol{y}^{(t)} - \max(\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)})\right)$, i.e. the log-magnitude DFT of two speakers is approximated by the element-wise maximum of their respective single-speech log-magnitude DFT $\boldsymbol{y}^{(t)} \approx \max(\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)})$. This approximation is based on the sparse nature of speech in time-frequency representations where each bin of a speech mixture spectrogram is dominated by a single speaker. We obtain the pitch-conditional observation probability $p(\boldsymbol{y}|x_1, x_2)$ by marginalization over $\boldsymbol{s}_k$, i.e.

$$p(\boldsymbol{y}|x_1, x_2) = \int\int p(\boldsymbol{y}|\boldsymbol{s}_1, \boldsymbol{s}_2) p(\boldsymbol{s}_1|x_1) p(\boldsymbol{s}_2|x_2) d\boldsymbol{s}_1 d\boldsymbol{s}_2. \quad (2)$$

This can be solved in closed form using single-speaker GMMs $p(\boldsymbol{s}_k|x_k)$ and the MIXMAX interaction model, i.e.

$$p(\boldsymbol{y}|x_1, x_2) = \sum_{m_1=1}^{M_{1,x_1}} \sum_{m_2=1}^{M_{2,x_2}} \alpha_{1,x_1}^{m_1} \alpha_{2,x_2}^{m_2} \quad (3)$$

$$\cdot \prod_{d=1}^{D} \left\{\mathcal{N}(y_d|\theta_{1,x_1}^{m_1,d}) \Phi(y_d|\theta_{2,x_2}^{m_2,d}) + \Phi(y_d|\theta_{1,x_1}^{m_1,d}) \mathcal{N}(y_d|\theta_{2,x_2}^{m_2,d})\right\},$$

where y_d denotes the d^{th} element of $\boldsymbol{y}$, $\theta_{k,x_k}^{m_k,d}$ gives the d^{th} element of the corresponding mean and variance of the single-speaker model of speaker k, and $\Phi(y|\theta) := \int_{-\infty}^{y} \mathcal{N}(x|\theta) dx$ represents the univariate cumulative normal distribution (details are in [1, 9]).

We perform model training using a set of pitch-labeled single-speaker utterances either in a speaker dependent (SD) or in a speaker independent (SI) fashion. For SD training only speaker specific speech utterances are used, whereas for SI training utterances from a large amount of different speakers are required. The transition probabilities, $p(x_k^{(t)}|x_k^{(t-1)})$ and prior distribution $p(x_k^{(1)})$, are obtained by maximum likelihood estimation and additional Laplace smoothing using the reference pitch values from the single-speaker recordings. The parameters of the single-speaker emission $p(\boldsymbol{s}_k^{(t)}|x_k^{(t)})$ are obtained by the EM-algorithm [15].

3. MODEL ADAPTATION

The availability of SD models is of great advantage, both in terms of accuracy as well as correct speaker assignment [1]. Even if we have SD models available, we might encounter different gain [16] or acoustic channel conditions in the test case, e.g. the spectral characteristics of each source signal might have changed due to multi-path propagation in a room. Model adaptation is useful to tune the available speaker models to the specific speaker characteristics and channel conditions that are present in a previously unseen recording. The aim is to adapt the model of each speaker involved given only the observed speech mixture.

3.1. Cepstrally Smoothed MLLR (csMLLR)

For the multi-pitch tracking framework, one individual GMM is used for each pitch state. The adaptation should not modify or destroy the harmonicity present in the model. Essentially, only the spectral envelopes of a speaker model should be subject to adaptation, while all fine-spectral structure modeled by each pitch-conditional GMM should remain unmodified. Hence, changing vocal tract characteristics and channel conditions can be captured, while still ensuring that each GMM represents its associated pitch. For this reason, we propose an affine transform of the log-spectrum mean vectors which is implicitly constrained in cepstral domain. Furthermore, we assume that the mean parameters of all GMMs associated with speaker model k are subject to the same transform. The transform $\tilde{\boldsymbol{T}}_k$ and $\tilde{\boldsymbol{b}}_k$ (bias vector) for the mean of speaker k, state x_k and component m_k is

$$\hat{\boldsymbol{\mu}}_{k,x_k}^{m_k} = \boldsymbol{W}\left(\tilde{\boldsymbol{T}}_k \boldsymbol{W} \boldsymbol{\mu}_{k,x_k}^{m_k} + \tilde{\boldsymbol{b}}_k\right) = \boldsymbol{W}\tilde{\boldsymbol{T}}_k \boldsymbol{W} \boldsymbol{\mu}_{k,x_k}^{m_k} + \boldsymbol{W}\tilde{\boldsymbol{b}}_k,$$

where matrix $\boldsymbol{W}$ denotes the (type I) $D \times D$ discrete cosine transform (DCT) matrix:

$$W_{i,j} = \begin{cases} \frac{1}{\sqrt{2D-2}} \cos\left(\frac{\pi}{D-1}(i-1)(j-1)\right) & \text{if } j \in \{1, D\}, \\ \frac{2}{\sqrt{2D-2}} \cos\left(\frac{\pi}{D-1}(i-1)(j-1)\right) & \text{otherwise,} \end{cases}$$

which essentially maps a mean vector $\boldsymbol{\mu}$ from log-spectral to cepstral domain.[5] The affine transform of the cepstral representation $\tilde{\boldsymbol{T}}_k \boldsymbol{W} \boldsymbol{\mu}_{k,x_k}^{m_k} + \tilde{\boldsymbol{b}}_k$ is back-transformed by $\boldsymbol{W}$ to log-spectral domain $\hat{\boldsymbol{\mu}}_{k,x_k}^{m_k}$.[6] We constrain the structure of the cepstral transform matrix to the form

$$\tilde{\boldsymbol{T}}_k = \left(\begin{array}{c|c} \boldsymbol{T}_k & \boldsymbol{0} \\ \hline \boldsymbol{0}^T & \boldsymbol{I} \end{array}\right),$$

where $\boldsymbol{T}_k$ denotes a $C \times C$ submatrix, $\boldsymbol{I}$ is the $(D-C) \times (D-C)$ identity matrix and $\boldsymbol{0}$ the $C \times (D-C)$ zero matrix. We constrain the bias vector $\tilde{\boldsymbol{b}}_k$ likewise. As a result, only the first C low-order coefficients of the cepstral representation of $\boldsymbol{\mu}$ are subject to the affine transform defined by $\boldsymbol{T}_k$ and $\boldsymbol{b}_k$. For $C = D$, no constraints are imposed on the transform, and the method produces equivalent results as MLLR. For small amounts of adaptation data, constraining the size of C can help to avoid overfitting.

3.2. EM Algorithm for csMLLR-Based FHMM Adaptation

We adapt parameters $\boldsymbol{T}_k$ and $\boldsymbol{b}_k$ by maximizing the log-likelihood under the observed speech mixture[7], i.e.

$$\begin{aligned} \text{LL}(\{\boldsymbol{T}_k\}, \{\boldsymbol{b}_k\}) &= \ln p(\mathcal{Y}|\{\boldsymbol{T}_k\}, \{\boldsymbol{b}_k\}) \quad (4) \\ &= \ln \sum_{\mathcal{X}} p(\mathcal{X}, \mathcal{Y}|\{\boldsymbol{T}_k\}, \{\boldsymbol{b}_k\}), \end{aligned}$$

[4] We omit the explicit dependence of random variables on t, where appropriate throughout the manuscript.

[5] Here, we do assume that $\boldsymbol{\mu}$ contains the bias bin at zero Hz, because it simplifies the application of the DCT transform and related notation.

[6] Matrix $\boldsymbol{W}$ is an involution, i.e. $\boldsymbol{W}\boldsymbol{W} = \boldsymbol{I}$.

[7] The remaining parameters of the model are omitted.

where

$$p(\mathcal{X},\mathcal{Y}|\{\boldsymbol{T}_k\},\{\boldsymbol{b}_k\}) = \prod_{k=1}^{K}\left[p(x_k^{(1)})\prod_{t=2}^{T}p(x_k^{(t)}|x_k^{(t-1)})\right]\prod_{t=1}^{T}p(\boldsymbol{y}^{(t)}|\{x_k^{(t)}\},\{\boldsymbol{T}_k\},\{\boldsymbol{b}_k\})$$

is the joint distribution of all observed data and hidden variables of an FHMM with K Markov chains. The distribution of the observation at one time instance given the hidden pitch states is (cf. (2))

$$\begin{aligned}p(\boldsymbol{y}^{(t)}|\{x_k^{(t)}\},\{\boldsymbol{T}_k\},\{\boldsymbol{b}_k\}) &= \int\!\cdots\!\int p(\boldsymbol{y}^{(t)}|\{\boldsymbol{s}_k^{(t)}\})\prod_{k=1}^{K}p(\boldsymbol{s}_k^{(t)}|x_k^{(t)},\boldsymbol{T}_k,\boldsymbol{b}_k)d\boldsymbol{s}_1^{(t)}\cdots d\boldsymbol{s}_K^{(t)}\\ &= \sum_{\{m_k\}}\prod_{k=1}^{K}\alpha_{k,x_k}^{m_k}\int\!\cdots\!\int p(\boldsymbol{y}^{(t)}|\{\boldsymbol{s}_k^{(t)}\})\\ &\quad\times\prod_{k=1}^{K}p(\boldsymbol{s}_k^{(t)}|x_k^{(t)},m_k,\boldsymbol{T}_k,\boldsymbol{b}_k)d\boldsymbol{s}_1^{(t)}\cdots d\boldsymbol{s}_K^{(t)},\end{aligned}$$

where

$$p(\boldsymbol{s}_k^{(t)}|x_k^{(t)},m_k,\boldsymbol{T}_k,\boldsymbol{b}_k) = \mathcal{N}(\boldsymbol{s}_k^{(t)}|\boldsymbol{W}\tilde{\boldsymbol{T}}_k\boldsymbol{W}\boldsymbol{\mu}_{k,x_k^{(t)}}^{m_k}+\boldsymbol{W}\tilde{\boldsymbol{b}}_k,\boldsymbol{\Sigma}_{k,x_k^{(t)}}^{m_k}).$$

It is difficult to maximize the log-likelihood in (4) directly. Instead, Jensen's inequality is applied to construct a lower bound on (4), which is in general easier to optimize [17]. For any distribution $q(\cdot)$, and any joint probability $p(\mathcal{X},\mathcal{Y})$, it follows from Jensen's inequality that

$$\ln\sum_{\mathcal{X}}p(\mathcal{X},\mathcal{Y}) = \ln\sum_{\mathcal{X}}q(\mathcal{X})\frac{p(\mathcal{X},\mathcal{Y})}{q(\mathcal{X})} \geq \sum_{\mathcal{X}}q(\mathcal{X})\ln\frac{p(\mathcal{X},\mathcal{Y})}{q(\mathcal{X})},$$

and equality holds if and only if $q(\mathcal{X}) = p(\mathcal{X}|\mathcal{Y})$. We systematically apply Jensen's inequality to construct the following variational lower bound on the $\mathrm{LL}(\{\boldsymbol{T}_k\},\{\boldsymbol{b}_k\})$ in (4):

$$\begin{aligned}\mathrm{LL} &\geq \text{const} + \sum_{\mathcal{X}}q(\mathcal{X})\ln p(\mathcal{X},\mathcal{Y}|\{\boldsymbol{T}_k\},\{\boldsymbol{b}_k\})\\ &\geq \text{const} + \sum_{\mathcal{X}}q(\mathcal{X})\sum_{t=1}^{T}\sum_{\{m_k\}}q(\{m_k\})\\ &\quad\times\ln\int\!\cdots\!\int p(\boldsymbol{y}^{(t)}|\{\boldsymbol{s}_k^{(t)}\})\prod_{k=1}^{K}p(\boldsymbol{s}_k^{(t)}|x_k^{(t)},m_k,\boldsymbol{T}_k,\boldsymbol{b}_k)d\boldsymbol{s}_1^{(t)}\cdots d\boldsymbol{s}_K^{(t)}\\ &\geq \text{const} + \sum_{\mathcal{X}}q(\mathcal{X})\sum_{t=1}^{T}\sum_{\{m_k\}}q(\{m_k\})\\ &\quad\times\int\!\cdots\!\int q(\{\boldsymbol{s}_k^{(t)}\})\sum_{k=1}^{K}\ln p(\boldsymbol{s}_k^{(t)}|x_k^{(t)},m_k,\boldsymbol{T}_k,\boldsymbol{b}_k)d\boldsymbol{s}_1^{(t)}\cdots d\boldsymbol{s}_K^{(t)},\end{aligned} \tag{5}$$

where const refers to all terms independent of $\{\boldsymbol{T}_k\}$ and $\{\boldsymbol{b}_k\}$. This lower bound is valid for an arbitrary choice of the variational distributions $q(\mathcal{X})$, $q(\{m_k\})$ and $q(\{\boldsymbol{s}_k^{(t)}\})$. Starting with an initial guess for the adaptation parameters, a local maximum of (4) can be found using the EM algorithm.

E-Step: The variational distributions are set such that the lower bound is tight[8] at the current parameter estimate, i.e.

$$q(\{x_k^{(t)}\}) = \sum_{\mathcal{X}\setminus\{x_k^{(t)}\}}p(\mathcal{X}|\mathcal{Y},\{\boldsymbol{T}_k^{(old)}\},\{\boldsymbol{b}_k^{(old)}\}), \tag{6}$$

$$q(\{m_k\}) = p(\{m_k\}|\boldsymbol{y}^{(t)},\{x_k^{(t)}\},\{\boldsymbol{T}_k^{(old)}\},\{\boldsymbol{b}_k^{(old)}\}),\text{and} \tag{7}$$

$$q(\{\boldsymbol{s}_k^{(t)}\}) = p(\{\boldsymbol{s}_k^{(t)}\}|\boldsymbol{y}^{(t)},\{x_k^{(t)}\},\{m_k\},\{\boldsymbol{T}_k^{(old)}\},\{\boldsymbol{b}_k^{(old)}\}). \tag{8}$$

Note that the calculation of Equations (6) and (7) is equivalent to the E-Step for exact parameter learning in FHMMs [18]. Specifically, (6) represents the marginal posterior, which can be obtained using the forward-backward algorithm of FHMMs as proposed in [18].

M-Step: The lower bound of the LL in (5) is maximized with respect to the parameters. For each speaker k, the parameters are updated according to $\{\boldsymbol{T}_k,\boldsymbol{b}_k\} = \arg\max_{\{\boldsymbol{T},\boldsymbol{b}\}}Q_k(\boldsymbol{T},\boldsymbol{b})$. The auxiliary function $Q_k(\boldsymbol{T}_k,\boldsymbol{b}_k)$ for speaker k is

$$Q_k(\boldsymbol{T}_k,\boldsymbol{b}_k) = \sum_{t,\{x_k\},\{m_k\}}\gamma_{t,\{x_k\},\{m_k\}}\,\mathrm{E}_{\{\boldsymbol{s}_k^{(t)}\}}\Big\{\ln\mathcal{N}(\boldsymbol{s}_k^{(t)}|\boldsymbol{W}\tilde{\boldsymbol{T}}_k\boldsymbol{W}\boldsymbol{\mu}_{k,x_k}^{m_k}+\boldsymbol{W}\tilde{\boldsymbol{b}}_k,\boldsymbol{\Sigma}_{k,x_k}^{m_k})\Big\}, \tag{9}$$

where we introduced the shorthand

$$\begin{aligned}\gamma_{t,\{x_k\},\{m_k\}} &= p(\{x_k^{(t)}\},\{m_k\},|\mathcal{Y},\{\boldsymbol{T}_k^{(old)}\},\{\boldsymbol{b}_k^{(old)}\})\\ &= q(\{x_k^{(t)}\})q(\{m_k\})\end{aligned}$$

to denote the posterior of states and components obtained in the previous E-Step (Equations (6) and (7)). The auxiliary function $Q_k(\boldsymbol{T}_k,\boldsymbol{b}_k)$ is obtained by plugging Equations (6), (7) and (8) into (5). The probability of the hidden single-speaker spectrum $\boldsymbol{s}_k^{(t)}$ has been replaced by its conditional expected value where the expectation $\mathrm{E}_{\{\boldsymbol{s}_k^{(t)}\}}\{\cdot\}$ is with respect to the distribution in (8). During the M-Step, the adaptation parameters can be optimized independently for each speaker.

As $Q_k(\cdot,\cdot)$ is jointly concave in $\boldsymbol{T}_k$ and $\boldsymbol{b}_k$, a global optimum can be obtained by setting the derivative to zero [19]. This leads to two cases: (i) $\frac{\partial Q_k(\boldsymbol{T}_k,\boldsymbol{b}_k)}{\partial\boldsymbol{T}_k} = 0$; assuming that $\boldsymbol{b}_k$ is fixed; (ii) $\frac{\partial Q_k(\boldsymbol{T}_k,\boldsymbol{b}_k)}{\partial\boldsymbol{b}_k} = 0$; assuming that $\boldsymbol{T}_k$ is fixed. The derivative of (i) and (ii) leads to a closed form solution for $\boldsymbol{T}$ and $\boldsymbol{b}$, respectively. Due to page limitations, we refer the reader for a detailed derivation to [20, 12]. Both (i) and (ii) are applied iteratively during the M-Step.

During the E-Step, the unknown single-speaker spectrum of every speaker is inferred, based on the currently available speaker models. During the M-Step, the expected single-speaker spectrum is used as a surrogate to the true single-speaker spectrum, and model parameters $\boldsymbol{T}_k$ and $\boldsymbol{b}_k$ are updated according to csMLLR. Unfortunately, the forward-backward algorithm as well as the calculation of sufficient statistics during the E-Step of the exact algorithm are intractable. Therefore, we make use of the fast pruning scheme developed for the MIXMAX interaction model [20].

4. EXPERIMENTS

The tracking performance of the algorithm is measured using E_{Total} [1], which is a slight extension of the error measure proposed

[8] Up to a term that does not depend on adaptation parameters.

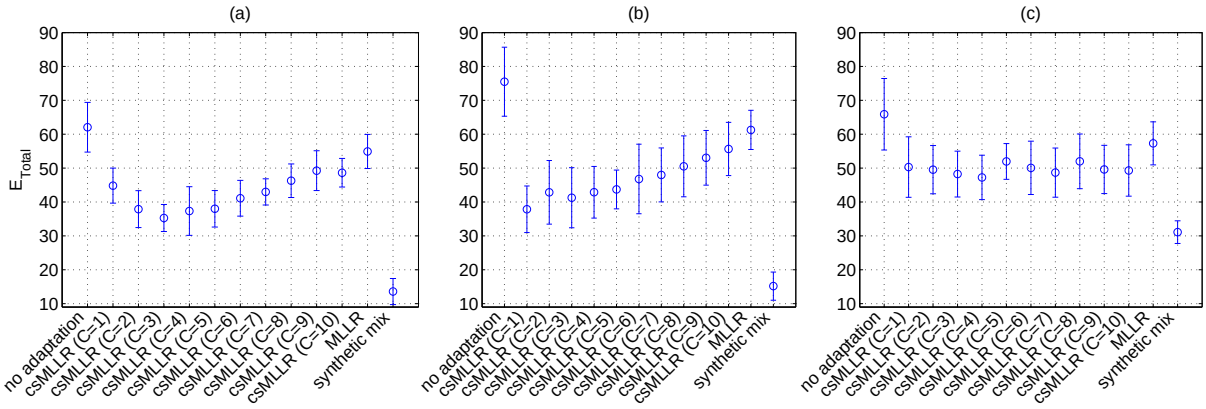

Fig. 2. Multi-pitch tracking performance in terms of E_{Total} after self-adaptation of SD models on real recordings. (a) male-female mixtures. (b) female-female mixtures. (c) male-male mixtures. 9 test mixtures were used per speaker pair, and error bars indicate the corresponding mean and standard deviation of E_{Total} for various methods. 'no adaptation': SD models of both speakers were used without adaptation. 'csMLLR': SD models of both speakers were adapted on the test mixture using csMLLR. 'MLLR': SD models of both speakers were adapted on the test mixture using MLLR. 'synthetic mix': SD models of both speakers were used without adaptation and applied for multi-pitch tracking on the synthetic mixture (i.e. no recording in room environment).

in [13] to additionally measure the influence of speaker assignment errors E_{Perm}, i.e. $E_{Total} = E_{01} + E_{02} + E_{10} + E_{12} + E_{20} + E_{21} + E_{Gross} + E_{Fine} + E_{Perm}$. E_{ij} denotes the percentage of time frames, where i pitch points are misclassified as j pitch points. E_{Perm} measures the percentage of frames, where the voicing decision is correct, but the pitch values are not assigned to the correct speakers. E_{Gross} is the percentage of frames where the voicing decision is correct and no permutation error has occurred, but at least one detected pitch value deviates more than 20% from the reference pitch value. E_{Fine} is defined as $E^{(1)}_{Fine} + E^{(2)}_{Fine}$, where $E^{(i)}_{Fine}$ denotes the frequency deviation in percent for speaker i, averaged over frames where no voicing, no gross and no permutation errors have occurred.

We use the proposed framework to perform self-adaptation on mixtures recorded in a real office environment where the spectral characteristics of each source signal have changed due to multi-path propagation or a different microphone transfer function. For this experiment, we used a total of 27 test mixtures, consisting of 9 test mixtures from three speaker pairs (female-female, male-female, male-male) from the GRID corpus [21], played through Yamaha MSP5A loudspeakers. The recording room has the dimensions $6.02 \times 5.32 \times 3\,\mathrm{m}$. One of the walls of the room has a large window, and the floor is covered with a carpet. The measured reverberation time (RT_{60}) was $\mathrm{RT}_{60} \approx 500$ ms; no particular effort was made to reduce the reverberation. For each recorded speech mixture, two GRID utterances were played back simultaneously with two loudspeakers positioned at different locations around a circular microphone array (with $0.15\,\mathrm{m}$ diameter). We process the recordings of one channel of this array. The distance between loudspeakers and the microphone was about $2\,\mathrm{m}$.

For each test speaker, 450 sentences from the GRID corpus were used to train SD GMMs. The reference pitch trajectories needed for training and evaluation were obtained using the RAPT method [22].[9] The observed features $\boldsymbol{y}^{(t)}$ are based on the log-spectrogram of the speech mixtures. The spectrogram is computed via the 1024 point DFT, using a Hamming window of length 32ms and step size of 10ms.[10] The sampling frequency is $f_s = 16$kHz. Each observation vector $\boldsymbol{y}^{(t)} \in \mathbb{R}^{64}$ is obtained by taking the log-magnitude of spectral bins 1-64, which corresponds to a frequency range up to 1000Hz.

For each test mixture, we applied self-adaptation using either csMLLR or MLLR and evaluated the resulting multi-pitch tracking performance. A summary of the results is shown in Figure 2, where the performance is additionally compared to the case where (i) no adaptation is performed and (ii) multi-pitch tracking is performed on the equivalent synthetic test mixture.[11] SD models without adaptation work very well when applied to a synthetic mixture, but result in heavily degraded performance when applied to the recorded mixture. Self-adaptation is able to improve the performance. A low value of C works better, as only few data for adaptation is available. Self-adaptation works best for mixtures of a male and female speaker, using csMLLR with $C = 3$.

5. CONCLUSIONS

We developed a model adaptation framework for FHMMs based on the EM algorithm and the MLLR technique to compensate any mismatch between training and testing conditions. We are able to adapt our models to a novel acoustic environment using only speech mixtures. Additionally, we propose a modification of the MLLR technique, where the adaptation of model parameters is constrained to modifications of the spectral envelope. This is beneficial in cases of few adaptation data. All developed methods are empirically compared for multipitch tracking. In future, we aim to adapt SI models in this self-adaptation scenario. We plan to extend our MLLR-based adaptation framework to additionally adapt the covariances of the speaker models. Furthermore, we plan to apply the proposed framework on other applications such as speech recognition, source separation and speaker identification.

[9] An implementation of the RAPT algorithm is provided by the Entropic speech processing system (ESPS) labeled as "get_f0" method.

[10] For the task of pitch estimation, the analysis window commonly includes at least 2-3 periods of the corresponding candidate f_0.

[11] For each recorded test mixture, a corresponding synthetic test mixture was created by linear superposition of the time-aligned original GRID utterances.

6. REFERENCES

[1] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 799 –810, 2011.

[2] Martin Cooke, John R. Hershey, and Steven J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, vol. 24, pp. 1–15, 2010.

[3] S.J. Rennie, J.R. Hershey, and P.A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 66 –80, 2010.

[4] J.R. Hershey, S. J. Rennie, P.A. Olsen, and T.T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer, Speech & Language*, vol. 24, pp. 45–66, 2010.

[5] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[6] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249 – 264, 1996.

[7] J.-L. Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[8] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695 –707, 2000.

[9] A. Nadas, D. Nahamoo, and M.A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

[10] R.C. Rose, E.M. Hofstetter, and D.A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994.

[11] R.J. Weiss and D. Ellis, "A variational EM algorithm for learning eigenvoice parameters in mixed signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 113 –116.

[12] M. Wohlmayr, *Probabilistic Model-Based Multiple Pitch Tracking of Speech*, Ph.D. thesis, Graz University of Technology, 2012.

[13] M. Wu, D. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[14] M.G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool, 2008.

[15] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. B30, pp. 1–38, 1977.

[16] Michael Wohlmayr and Franz Pernkopf, "EM-based gain adaptation for probabilistic multipitch tracking," in *International Conference on Spoken Language Processing (Interspeech)*, 2011, pp. 1969–1972.

[17] C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[18] Z. Ghahramani and M.I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[19] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[20] M. Wohlmayr and F. Pernkopf, "Model-based multiple pitch tracking using factorial HMMs: Model adaptation and inference," *IEEE Transactions on Audio, Speech, and Language Processing*, accepted for publication.

[21] M.P. Cooke, J. Barker, S.P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2005.

[22] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Kleijn W.B. and Paliwal K.K. [Ed], Speech Coding and Synthesis, Elsevier Science*, pp. 495–518, 1995.