A MIXTURE MAXIMIZATION APPROACH TO MULTIPITCH TRACKING WITH FACTORIAL HIDDEN MARKOV MODELS

M. Wohlmayr, M. Stark, F. Pernkopf

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

ABSTRACT

We present a simple and efficient feature modeling approach for tracking the pitch of two speakers speaking simultaneously. We model the spectrogram features of single speakers using Gaussian mixture models in combination with the minimum description length model selection criterion. Furthermore, the mixture maximization (MIXMAX) interaction model is employed to yield a probabilistic representation for the mixture of both speakers. Finally, a factorial hidden Markov model is applied for tracking. We demonstrate experimental results on two databases, and show the excellent performance of the proposed method in comparison to a well known multipitch tracking algorithm based on correlogram features.

Index Terms— Multipitch tracking, factorial hidden Markov model, mixture maximization, Gaussian mixture model.

1. INTRODUCTION

Estimation and tracking of pitch is important for many algorithms and applications in speech and audio signal processing, e.g. single-channel blind source separation [1], speech compression, and prosodic speech analysis. While well performing algorithms do exist for the case of a single speaker in a clean recording [2], the same task of pitch estimation is more difficult for noisy speech and multiple speakers talking simultaneously.

Wu et al. [3] proposed an approach for robust multipitch tracking. It is based on the unitary model of pitch perception [4], 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. In [5] and [6] a factorial hidden Markov model (FHMM) [7] is used, which provides the natural means for tracking the pitch trajectories of multiple speakers. Recently, we proposed to model the spectrogram features of speech mixtures with Gaussian mixture models (GMMs) [8], resulting in a significant performance improvement over [3]. Specifically, the usage of speaker dependent models enabled to correctly link the pitch estimates to their corresponding speakers.

In this paper, we introduce an advanced feature modeling approach for multipitch tracking. In particular, pitch-dependent spectrogram features of single speakers are modeled with Gaussian mixture models (GMMs), whereas the minimum description length (MDL) [9, 10] criterion is applied to find the optimal number of Gaussian components. Using the mixture maximization (MIXMAX) [11] approach, the individual single speaker GMMs are then combined to yield a probabilistic model of the speech mixture, conditioned on the pitch states of both speakers. This statistical model is used within the framework of FHMMs. We refer to this method as MIXMAX-FHMM. In the experiments, we compare our MIXMAX-FHMM approach to a well known method proposed by Wu et al. [3] based on correlogram features and using an HMM for tracking. We refer to this method as COR-HMM. We compare both methods on two different databases, namely the Mocha-TIMIT [12] and the GRID database [13]. For both databases, our method significantly outperforms the correlogram-based method for speech utterances mixed at 0dB. Further, we demonstrate the excellent performance on noise corrupted mixtures from Mocha-TIMIT.

The paper is organized as follows: Section 2 introduces FHMMs for multipitch tracking. In Section 3, we motivate our feature model based on the MIXMAX approach. Section 4 shortly reviews methods for tracking in FHMMs. The experimental results are discussed in Section 5. Section 6 concludes the paper.

2. FACTORIAL HIDDEN MARKOV MODELS

Factorial hidden Markov models (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. For simplicity, we present the case of two Markov chains depicted as factor graph in Fig. 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, realizations of the observed random variables at t are collected in a vector $\boldsymbol{y}^{(t)} \in \mathbb{R}^D$. Each $x_k^{(t)}$ represents a discrete random variable, where for simplicity all variables are assumed to have cardinality $|X|$. The edges between nodes indicate a direct conditional dependency between 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 $\boldsymbol{y}^{(t)}$ on the hidden variables of the same time frame are defined by the observation probability $p(\boldsymbol{y}^{(t)}|x_1^{(t)}, x_2^{(t)})$. Finally, the prior distribution of the hidden variables in every chain is denoted by $p(x_k^{(1)})$. Denoting the whole sequence of variables, i.e. $\{x^{(t)}\} = \bigcup_{t=1}^{T}\{x_1^{(t)}, x_2^{(t)}\}$ and $\{\boldsymbol{y}^{(t)}\} = \bigcup_{t=1}^{T}\boldsymbol{y}^{(t)}$, the joint distribution of all variables is given by

$$p(\{x^{(t)}\}, \{\boldsymbol{y}^{(t)}\}) = p(\{x^{(t)}\})p(\{\boldsymbol{y}^{(t)}\}|\{x^{(t)}\}) = \prod_{k=1}^{2}\left[p(x_k^{(1)})\prod_{t=2}^{T}p(x_k^{(t)}|x_k^{(t-1)})\right]\prod_{t=1}^{T}p(\boldsymbol{y}^{(t)}|x_1^{(t)}, x_2^{(t)}).$$

This work was supported by the Austria Science Fund (project number P19737-N15).

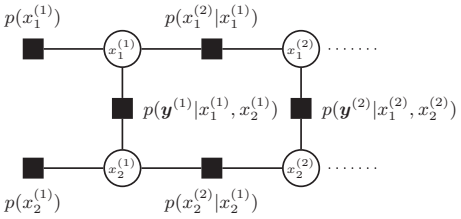

Fig. 1. A factorial HMM shown as a factor graph [14]. Factor nodes are depicted as shaded rectangles together with their functional description. Hidden variable nodes are shown as circles. Here, observed variables $\boldsymbol{y}^{(t)}$ are absorbed into factor nodes.

The number of possible hidden states per time frame is $|X|^2$. As pointed out in [7], this could also be accomplished by an ordinary HMM. The main difference, however, is the constraint placed upon the transition structure. While an HMM with $|X|^2$ states would allow any $|X|^2 \times |X|^2$ transition matrix between two hidden states, the FHMM is restricted to two $|X| \times |X|$ transition matrices.

2.1. FHMM parameter settings

As in most previous work for multipitch tracking [3, 5, 6], we restrict ourselves to two simultaneously speaking subjects, i.e. two Markov chains. Each Markov chain models the pitch trajectory of one speaker, hence the hidden variable $x_k^{(t)}$ denotes the pitch state of speaker k at time t. Each hidden variable has $|X| = 170$ states, where state value '1' refers to 'no pitch' (i.e. unvoiced or silent), and state values '2'-'170' encode different pitch frequencies ranging from 80 to 500Hz. Specifically, the pitch value f_0 corresponding to state $x > 1$ is obtained as $f_0 = \frac{f_s}{30+x}$, where sampling rate $f_s = 16$kHz. Similar to [3], this results in a nonuniform quantization of the pitch interval, where low pitch values have a more fine grained resolution than high pitch values.

3. MIXMAX INTERACTION MODEL

At each time frame t, the FHMM models the observed log-spectrum $\boldsymbol{y}^{(t)}$ of the mixture signal by the observation probability $p(\boldsymbol{y}^{(t)}|x_1^{(t)}, x_2^{(t)})$. Recently, we modelled the spectrogram features for each pitch pair with one individual GMM [8]. In this work, however, the design of $p(\boldsymbol{y}^{(t)}|x_1^{(t)}, x_2^{(t)})$ is guided by the insight that the log-spectrogram of two speakers can be approximated by their elementwise maximum [11]. Specifically, for each time instant t, $\boldsymbol{y}^{(t)} \approx \max(\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)})$, where $\boldsymbol{s}_i^{(t)}$ is the log-spectrum of speaker i. Thus, we might think of $\boldsymbol{y}^{(t)}$ being generated by the stochastic model in Fig. 2. For a fixed pitch value related to state $x_i^{(t)}$, speaker i generates a log-spectrum, $\boldsymbol{s}_i^{(t)}$, that is randomly drawn from the *single speaker model* $p(\boldsymbol{s}_i^{(t)}|x_i^{(t)})$. Both log-spectra are then combined via the elementwise maximum operator to form the observable log-spectrum $\boldsymbol{y}^{(t)}$. Thus, $p(\boldsymbol{y}^{(t)}|\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)}) = \delta\left(\boldsymbol{y}^{(t)} - \max(\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)})\right)$, where $\delta(\cdot)$ denotes the Dirac delta.

In general, we obtain the observation probability by marginalizing over the unknown single speaker log-spectra:

$$p(\boldsymbol{y}^{(t)}|x_1^{(t)}, x_2^{(t)}) = \int\int p(\boldsymbol{y}^{(t)}|\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)}) p(\boldsymbol{s}_1^{(t)}|x_1^{(t)}) p(\boldsymbol{s}_2^{(t)}|x_2^{(t)}) d\boldsymbol{s}_1^{(t)} d\boldsymbol{s}_2^{(t)}. \quad (1)$$

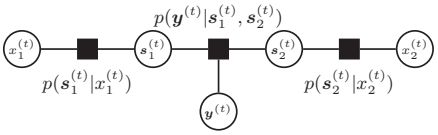

Fig. 2. Pitch dependent generation of the mixture log-spectrum $\boldsymbol{y}^{(t)}$. Both speakers produce a log-spectrum $\boldsymbol{s}_i^{(t)}$ in dependency on pitch state $x_i^{(t)}$. The observed log-spectrum $\boldsymbol{y}^{(t)}$ of the speech mixture is approximated by the elementwise maximum of both single speaker log-spectra.

For the sake of brevity, we omit the explicit dependence of random variables on t, where appropriate. We use GMMs to model the state-conditional single speaker spectra of both speakers, $i \in \{1, 2\}$, according to

$$p(\boldsymbol{s}_i|x_i) = p(\boldsymbol{s}_i|\boldsymbol{\Theta}_{i,x_i}) = \sum_{m=1}^{M_{i,x_i}} \alpha_{i,x_i}^m \mathcal{N}\left(\boldsymbol{s}_i|\boldsymbol{\theta}_{i,x_i}^m\right), \quad (2)$$

where $M_{i,x_i} \geq 1$ is the number of mixture components, and α_{i,x_i}^m corresponds to the weight of each component $m = 1, \ldots, M_{i,x_i}$. These weights are constrained to be positive, $\alpha_{i,x_i}^m \geq 0$, and $\sum_{m=1}^{M_{i,x_i}} \alpha_{i,x_i}^m = 1$. The corresponding GMM is fully specified by the parameter set $\boldsymbol{\Theta}_{i,x_i} = \left\{\alpha_{i,x_i}^m, \boldsymbol{\theta}_{i,x_i}^m\right\}_{m=1}^{M_{i,x_i}}$, where $\boldsymbol{\theta}_{i,x_i}^m = \left\{\boldsymbol{\mu}_{i,x_i}^m, \boldsymbol{\Sigma}_{i,x_i}^m\right\}$. Furthermore, we assume diagonal covariance matrices. Hence, by introducing speaker specific GMMs and marginalizing over $\boldsymbol{s}_i$, we obtain for (1):

$$p(\boldsymbol{y}|x_1, x_2) = \sum_{m=1}^{M_{1,x_1}} \sum_{n=1}^{M_{2,x_2}} \alpha_{1,x_1}^m \alpha_{2,x_2}^n \prod_{d=1}^{D} \left\{\mathcal{N}(y_d|\theta_{1,x_1}^{m,d}) \Phi(y_d|\theta_{2,x_2}^{n,d}) + \Phi(y_d|\theta_{1,x_1}^{m,d}) \mathcal{N}(y_d|\theta_{2,x_2}^{n,d})\right\},$$

where y_d gives the d^{th} element of $\boldsymbol{y}$, $\theta_{i,x_i}^{m,d}$ gives the d^{th} element of the corresponding mean and variance, and $\Phi(y|\theta) = \int_{-\infty}^{y} \mathcal{N}(x|\theta) dx$ denotes the univariate cumulative normal distribution. Given a set of N_i single speaker log-spectra for speaker i, $\mathcal{S}_i = \{\boldsymbol{s}_i^{(1)}, ..., \boldsymbol{s}_i^{(N_i)}\}$, together with corresponding reference pitch labels, $\{x_i^{(1)}, ..., x_i^{(N_i)}\}$ we can easily learn a speaker dependent GMM $p\left(\boldsymbol{s}_i|\boldsymbol{\Theta}_{i,x_i}\right)$ for each pitch state x_i, and each speaker i, using the EM algorithm [15]. Accordingly, we have to determine 170 GMMs for each speaker, i.e. one GMM for each pitch state x_i. Further, we use MDL [9, 10] to determine the number of components of each GMM automatically.

In principle, this approach can be extended to more than two speakers, however resulting in an increased computational complexity.

4. TRACKING

Given the set of observations $\{\boldsymbol{y}^{(t)}\}$, the task of tracking involves searching the sequence of hidden states $\{x^{(t)}\}^*$ that maximizes the conditional distribution $p(\{x^{(t)}\}|\{\boldsymbol{y}^{(t)}\})$. For HMMs, the exact solution to this problem is found by the Viterbi algorithm. For FHMMs, an exact solution can be determined using the junction tree algorithm [16], however its computational complexity increases exponentially with the number of hidden Markov chains. Several algorithms to obtain approximate solutions for the sake of reduced

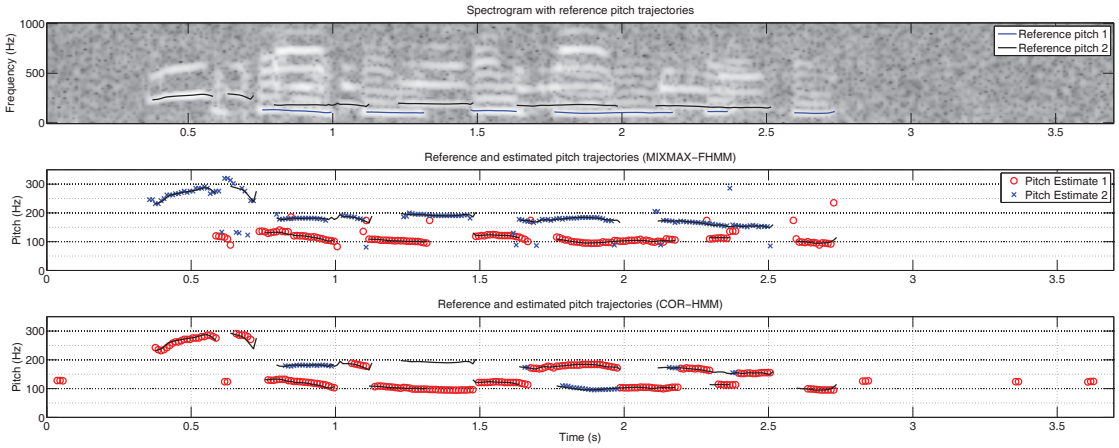

Fig. 3. Trajectories found by MIXMAX-FHMM and COR-HMM, applied on a test mixture from Mocha-TIMIT that was additionally corrupted with white noise at 10dB SNR. The top panel shows the spectrogram of the noisy speech mixture, together with both reference pitch trajectories. The middle and bottom panel show the reference pitch, together with the pitch estimates resulting from the MIXMAX-FHMM and COR-HMM method, respectively.

complexity are derived in [7] from the framework of variational inference. The sum-product algorithm [14] can be derived under a similar setting of variational principles [17], although more intuitive derivations exist for graphs without loops. When applied on a graph with loops, as in the case of an FHMM, the solutions are in general not guaranteed to converge and can only approximate the optimal solution. In this work, we use the max-sum algorithm (a variant of the sum-product algorithm) for tracking. In [6], experimental results suggested that the obtained solutions sufficiently approximate the exact solution, while computational complexity is much lower.

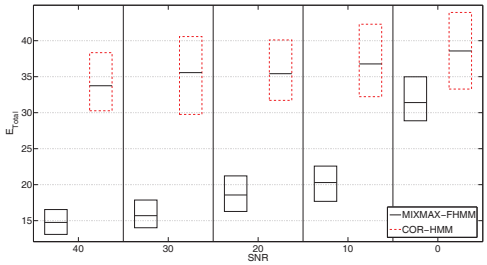

Fig. 4. Performance of MIXMAX-FHMM and COR-HMM on test instances of Mocha-TIMIT corrupted with white Gaussian noise at different SNR levels. Each box depicts the median as well as the lower and upper quartile of E_{Total} on the test set.

5. EXPERIMENTAL RESULTS

We compare the performance of our approach (MIXMAX-FHMM) to the COR-HMM [3] method on two databases. First, the Mocha-TIMIT database [12] 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 reference pitch $f_0[t]$ was acquired using the RAPT method [2] together with manual removal of erroneous pitch estimates in nonaudible regions.[1] The speaker dependent GMMs were trained on 400 sentences each, while 60 test instances were obtained by mixing the remaining male and female utterances at 0dB. Second, two male and two female speakers were selected from the GRID database [13], where 500 English sentences are available per speaker. For each speaker, 497 sentences were used to train speaker dependent GMMs, while the remaining three sentences were used for testing. Test mixtures were created for each speaker pair, including same-gender mixtures, resulting in a total of 54 test mixtures (9 mixtures for each of the 6 speaker pairs). As no laryngograph signals are available for this database, the reference pitch trajectories were obtained directly from the single speech utterances using the RAPT method.

The input features $\boldsymbol{y}^{(t)}$ of the proposed method are based on the log-spectrogram of the speech mixture. Given an input signal at sampling rate $f_s = 16$kHz, we compute the spectrogram via the 1024 point FFT, using a Hamming window of length 32ms and step size of 10ms. Next, we obtain each observation vector $\boldsymbol{y}^{(t)} \in \mathbb{R}^{64}$ by taking the log of spectral bins 2-65, which corresponds to a frequency range up to 1000Hz. This covers the most relevant frequency range, while keeping the model complexity low.

Both transition matrices of the FHMM, $p(x_k^{(t)}|x_k^{(t-1)})$, are obtained by counting and normalizing the transitions of the reference pitch values from single speaker recordings in the training set. Additionally, we apply Laplace smoothing on both transition matrices. Prior distributions $p(x_k^{(1)})$ are obtained likewise.

For every test instance, each method estimates two pitch trajectories, $\tilde{f}_0^{(1)}[t]$ and $\tilde{f}_0^{(2)}[t]$. 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, $(\tilde{f}_0^{(1)} \rightarrow f_0^{(1)}, \tilde{f}_0^{(2)} \rightarrow f_0^{(2)})$ or $(\tilde{f}_0^{(1)} \rightarrow f_0^{(2)}, \tilde{f}_0^{(2)} \rightarrow f_0^{(1)})$, the one is chosen for which the overall quadratic error is smaller.

To evaluate the resulting estimates, we use the error measure proposed in [3]: 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 pitch points estimated whereas only one pitch is present. The pitch frequency deviation is defined as $\Delta f^{(k)}[t] = (|\tilde{f}_0^{(k)}[t] - f_0[t]|)/f_0[t]$, where $f_0[t]$ denotes the reference chosen for $\tilde{f}_0^{(k)}[t]$. The gross detection error rate E_{Gross} is the percentage of time frames where the algorithm correctly detects the presence of

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

		E_{01}	E_{02}	E_{10}	E_{12}	E_{20}	E_{21}	E_{Gross}	E_{Fine}	E_{Total}
MIXMAX-FHMM	Mean	3.62	0.21	0.96	3.50	0.02	1.48	1.35	2.87	**14.00**
	Std	1.45	0.27	0.63	1.48	0.09	0.89	1.02	0.35	2.21
COR-HMM [3]	Mean	5.61	0.10	3.66	1.44	0.40	7.26	12.39	2.68	**33.53**
	Std	2.30	0.23	1.33	1.25	0.49	2.70	5.33	1.66	8.19

Table 1. Error measure (in percent) of both methods on Mocha-TIMIT database.

		E_{01}	E_{02}	E_{10}	E_{12}	E_{20}	E_{21}	E_{Gross}	E_{Fine}	E_{Total}
MIXMAX-FHMM	Mean	0.89	0.00	6.66	2.13	1.82	12.33	1.16	2.81	**27.81**
	Std	1.26	0.00	3.15	1.84	2.62	6.29	1.86	0.79	10.96
COR-HMM	Mean	1.00	0.08	8.46	0.87	2.56	19.97	17.60	3.30	**53.83**
	Std	1.49	0.20	3.62	1.23	3.11	7.70	10.33	2.79	12.99

Table 2. Error measure (in percent) of both methods on GRID database.

one pitch point (two pitch points), 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_{Fine} is the average frequency deviation in percent at time frames where $\Delta f^{(k)}[t]$ is smaller than 20%. The overall error, E_{Total}, is defined as the sum of all error terms: $E_{Total} = E_{01} + E_{02} + E_{10} + E_{12} + E_{20} + E_{21} + E_{Gross} + E_{Fine}$. The performance of both methods on Mocha-TIMIT and GRID is shown in table 1 and 2, respectively. On both databases, MIXMAX-FHMM is superior in terms of accuracy, as well as correct assignment of pitch trajectories to their corresponding speakers. In particular, we are able to reduce E_{Total} by at least 48% relative.
In a second experiment, each test utterance from Mocha-TIMIT was mixed with white Gaussian noise at different SNR conditions, ranging from 40dB down to 0dB in 10dB steps. For each SNR condition, the performance of MIXMAX-FHMM was compared with COR-HMM, where the parameters of MIXMAX-FHMM remained optimized for clean speech. The resulting performance of both methods is shown in Fig. 4. Fig. 3 shows tracking results for one test mixture corrupted with white Gaussian noise at 10dB SNR.

6. CONCLUSIONS

We proposed an efficient feature modelling approach for multipitch tracking. We model the pitch-conditional log-spectrogram features of single speakers using GMMs, and combine these models using the MIXMAX approach to obtain a probabilistic model of the observed speech mixture. We demonstrated the performance of the proposed method on two different databases, and showed its superior performance over a well known multipitch tracking method [3] based on correlogram features. In particular, we were able to reduce E_{Total} by at least 48% relative.

References

[1] D.P. Morgan, E.B. George, L.T. Lee, and S.M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 407–424, 1997.

[2] D. Talkin, "A robust algorithm for pitch tracking," in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Eds., pp. 495–518. Elsevier Science, 1995.

[3] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 229–241, 2003.

[4] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, 1997.

[5] F.R. Bach and M.I. Jordan, "Discriminative training of hidden Markov models for multiple pitch tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 2005, pp. 489–492.

[6] M. Wohlmayr and F. Pernkopf, "Multipich tracking using a factorial hidden Markov model," in *Proceedings of Interspeech*, 2008.

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

[8] M. Wohlmayr and F. Pernkopf, "Finite mixture spectrogram modeling for multipitch tracking using a factorial hidden Markov model," in *Proceedings of Interspeech*, 2009.

[9] G.J. McLachlan and K.E. Basford, *Mixture Models*, Marcel Dekker, 1988.

[10] F. Pernkopf and D. Bouchaffra, "Genetic-based EM algorithm for learning Gaussian mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1344–1348, 2005.

[11] 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.

[12] A. Wrench, "A multichannel/multispeaker articulatory database for continuous speech recognition research," *Phonus*, vol. 5, pp. 3–17, 2000.

[13] 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, 2006.

[14] F. R. Kschischangand, B. J. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, pp. 498–519, 2001.

[15] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistic Society*, vol. 30, no. B, pp. 1–38, 1977.

[16] M.I. Jordan [Ed], *Learning in Graphical Models*, MIT Press, 1999.

[17] T.P. Minka, "Divergence measures and message passing," Tech. Rep., Microsoft Research Cambridge, 2005.