DISCRIMINATIVE TRAINING OF HIDDEN MARKOV MODELS FOR MULTIPLE PITCH TRACKING

Francis R. Bach

Computer Science Division
University of California
Berkeley, CA 94720, USA
fbach@cs.berkeley.edu

Michael I. Jordan

Computer Science and Statistics
University of California
Berkeley, CA 94720, USA
jordan@cs.berkeley.edu

ABSTRACT

We present a multiple pitch tracking algorithm that is based on direct probabilistic modeling of the spectrogram of the signal. The model is a factorial hidden Markov model whose parameters are learned discriminatively from the Keele pitch database [1]. Our algorithm can track several pitches and determines the number of pitches that are active at any given time. We present simulation results on mixtures of several speech signals and noise, showing the robustness of our approach.

1. INTRODUCTION

Pitch tracking is a fundamental problem in speech and music processing, and the design of robust algorithms for single or multiple pitch determination has been an active topic of research in acoustic signal processing [2, 3, 4, 5, 6, 7]. Most pitch extraction algorithms first build a set of nonlinear features (e.g., the correlogram or the cepstrum) that exhibit special behavior when voiced speech is uttered and then model this behavior to track pitch. In the presence of multiple voiced signals that mix additively, it is natural to consider modeling directly the signals or a linear representation thereof (such as the spectrogram) in order to preserve additivity and make it possible to use models for one pitch in order to extract multiple pitches. In this paper, we work with the magnitude of spectrogram. The magnitude is not a linear representation, but because of the sparsity of speech and music signals in the spectrogram, it can be well approximated as such [8].

Working directly with the spectrogram requires a detailed probabilistic model for characterizing pitch. In this paper, we consider a variant of a hidden Markov harmonic model and use the framework of graphical models to build the model, learn it from data and design efficient inference algorithms [9]. In particular, we use recent developments from the machine learning literature to capture the appropriate properties of speech and music; in particular, we make use of nonparametric priors to capture smoothness of the spectral envelope, and we improve extraction performance by using discriminative training of the models [10]. We present the graphical model in Section 2, the inference algorithm in Section 3 and the learning algorithm in Section 4. In Section 5, we test our algorithm on a variety of challenging pitch extraction tasks.

This work was supported by a grant from Intel Corporation, and a graduate fellowship to Francis Bach from Microsoft Research.

2. GRAPHICAL MODEL FOR PITCH EXTRACTION

In this paper, we assume that the speech signals are sampled at 5.5 KHz. Given a real one-dimensional signal x_t, $t = 1, \ldots, T$, the *spectrogram* s is defined as the short-time windowed Fourier transform of x; i.e., the signal x is cut into N overlapping frames of length M, and the spectrogram s is defined as the $N \times P$ matrix whose n-th column $s_n \in \mathbb{R}^P$ is the P-point FFT of a windowed version of the n-th frame.[1] In this paper, we model the magnitude of the spectrogram and refer to the magnitude of the spectrogram simply as the spectrogram. Since the speech signals are real, the FFT is symmetric and we only need to consider the first $P/2$ frequencies.

2.1. Additive model

The input to our pitch tracker is the sequence $s_n \in \mathbb{R}^P$, $n = 1, \ldots, N$, where N is the number of frames, equal to a constant times the duration T of the signal x. We use an additive model for the spectrogram, i.e., if K speakers are potentially present, we model the n-th frame as the superposition of K signals $u_n^k \in \mathbb{R}^P$ plus noise, i.e., $s_n = \sum_{k=1}^K u_n^k + \varepsilon_n$. Note that the acoustics are not additive for the magnitude of the spectrogram; however, since signals from two different speakers have small overlap [8], the linearity is a reasonable approximation. The advantage of using the magnitude is that the modeling of the smoothness of the spectral envelope is easily achieved using spline smoothing techniques, as described in Section 2.2.

2.2. Harmonic model

We use a harmonic model in the frequency domain, which amounts to modeling the spectrogram of voiced speech as an amplitude-modulated comb [3]. We model each speaker k at time frame n with four variables:

- *Voiced/unvoiced*: v_n^k is a binary variable which is equal to one if the speaker k utters voiced speech at time n, and equal to zero otherwise (either non-voiced speech or no speech uttered).
- *Pitch*: ω_n^k is the frequency of the pitch of speaker k at time n, scaled so that it is equal to the distance between two harmonic peaks in the spectrogram.

[1] In simulations, we use frames of length 40ms sampled every 10ms, a Hanning window and a 512-point FFT.

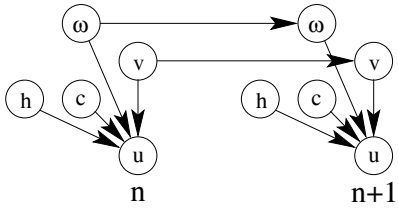

Fig. 1. HMM for one speaker for two time frames n and $n+1$ (time subscripts are omitted).

- *Harmonics*: h_n^k is a set of vectors of harmonic amplitudes if the speech is voiced. There is one harmonic vector $h_{n\omega}^k$ for each potential pitch value ω. The dimension of $h_{n\omega}^k$ is equal to $\lfloor P/2\omega \rfloor$. Given that the signal is non-voiced (i.e., given $v_n^k = 0$), then all sets of harmonic amplitudes for all ω are independent from all other variables, while given that the signal is voiced and given the pitch ω, the entire set $\{h^k n\omega', \omega' \neq \omega\}$ is independent from other variables.
- *Constant term*: c_n^k is the constant amplitude of non-voiced portions. Given $v_n^k = 1$, c_n^k is independent from all other variables.

The graphical model describing the model for a single speaker is a simple Hidden Markov model (HMM) and is shown in Figure 1. The conditional probability distributions that are needed to fully specify the model reflect the known psychoacoustics and statistical properties of pitch [11, 2, 3] and are as follows:

- $p(v_{n+1}^k | v_n^k)$ is a constant transition matrix T_v with four elements.
- $p(\omega_{n+1}^k | \omega_n^k)$: the pitch is discretized on a grid with $n_\omega = 300$ elements, and each logarithm of the row of the transition matrix is equal to (up to additive constants) $\alpha_1(\omega_{n+1}^k - \omega_n^k)^2 + \alpha_2\omega_{n+1}^k + \alpha_3(\omega_{n+1}^k)^{-1}$. Note that the high number of values for the discretization of the pitch frequency is necessary in order to obtain good pitch extraction performance.
- $p(h_{n\omega}^k)$: for each value of the pitch ω, $h_{n\omega}^k$ is modelled as the restriction of a smooth function on $[0, P/2]$—i.e., a function with bounded second derivative—to all multiples of ω. That is, $(h_{n\omega}^k)_i$ is equal to $g(i\omega)$, where g is a function such that $\int |g^{(2)}|^2$ is bounded. g is usually referred to as the *spectral envelope* [3].
  Following [12], $h_{n\omega}^k$ can thus be modelled as a Gaussian process on the line $[0, P/2]$ observed at multiples of the fundamental frequency ω; this implies that $h_{n\omega}^k$ can be written as $h_{n\omega}^k = K_\omega a_{n\omega}^k + T_\omega b_{n\omega}^k$, where K_ω is the "kernel matrix" defined as $(K_\omega)_{ij} = (\frac{1}{2}ij\min\{i,j\} - \frac{1}{6}\min\{i,j\}^3)\omega^3$, and T_ω is a matrix with two columns, one constant and one linear function of the frequency. The auxiliary variables $a_{n\omega}^k$ and $b_{n\omega}^k$ are normal with mean 0 and covariance matrices $(\alpha_4 K_\omega + \alpha_5 I)^{-1}$ and $\alpha_6 I$.
- The variable c_n^k is normal with mean α_4 and variance α_5.
- Observation model: given ω_n^k, h_n^k, c_n^k and v_n^k, the signal u_n^k is equal to $B(\omega_n^k)h_{n\omega_n^k}^k$ if $v_n^k = 1$, and equal to $c_n^k e$ if $v_n^k = 0$, where e is the constant vector of all ones. The i-th column of the matrix $B(\omega)$ is a bump centered at frequency $i\omega$, defined as the Fourier transform of the window. See Figure 2. Thus, voiced speech is modeled as a weighted sum of bumps at multiples of the fundamental frequency, where the amplitude of the bump extends to a smooth spectral envelope.

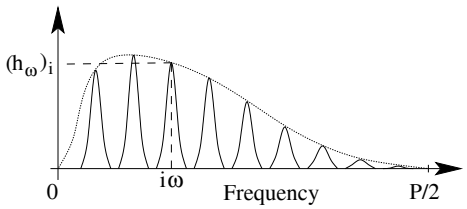

Fig. 2. Spectral envelope (dotted) and harmonic model (plain).

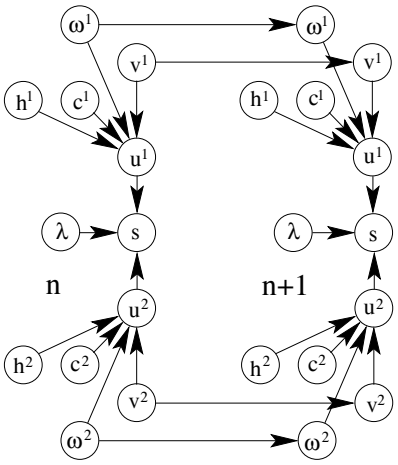

Fig. 3. Factorial HMM for two speakers for two time frames n and $n+1$ (time subscripts are omitted).

2.3. Factorial hidden Markov models

The K models for each speaker can be joined into a single graphical model, a "factorial HMM," where the $2K$ Markov chains evolve independently (see Figure 3 for the model with two speakers). The parameter λ_n is the variance of the Gaussian noise ε_n at time n. We assume it has a uniform distribution and it is discretize to an uniform logarithmic grid with $n_\lambda = 10$ elements.

2.4. Related models

Our graphical model resembles the models presented previously in [4], [5], [6] and [7]. In [4] and [5], the graphical model is defined on features rather than on the speech signal directly (or its spectrogram), which abandons the additive structure of the mixing and makes it more difficult to estimate several pitch tracks. In [6] and [7], harmonic models are used but most parameters are not learned from data, and the harmonic model does not include a smoothness prior which is crucial to avoid pitch halving. Also, models that are learned from data such as [4] or [5] use maximum likelihood training while we use discriminative training, which is more expensive but leads to superior performance (see Section 5).

3. PITCH EXTRACTION

In the following sections, we use the shorthand x to denote the set of variables $(x_n^k)_{k,n}$ for all k and n, while we use the shorthand x^k to denote the set of variables $(x_n^k)_n$ for all n. If we denote $z = (\omega, v, h, c, \lambda)$, then the task of inference is to compute, given some data s, $\arg\max_z p(z|s)$. Minimization with respect to (h, λ) can be done in closed form and thus we are left with the task of maximizing with respect to (ω, v).

3.1. One speaker

With one speaker, this is simply inference in an HMM where the hidden state has a number of values proportional to n_ω, and the complexity of inference for a speech of duration T is thus $O(Tn_\omega)$ for computing potentials and $O(Tn_\omega^2)$ for the Viterbi algorithm [13].

3.2. Two or more speakers

With m speakers, we have a factorial HMM with $2m$ uncoupled Markov chains with n_ω or 2 states each, thus the complexity of exact inference is $O(Tn_\omega^m)$ for computing potentials and $O(Tn_\omega^{m+1})$ for a structured Viterbi algorithm [13]. Given that searching of a space of size n_ω^2 is the most expensive we are willing to afford (since n_ω is large), we use the following approximate scheme which is a simple extension of similar schemes used for single pitch tracking (e.g., [7]):

1. Recursively estimate the m pitches by finding one single pitch track and subtracting the corresponding estimated harmonic signals.
2. Construct a pool of p_ω pitch value candidates for each time step, by storing local minima in the m Viterbi algorithms of step 1.
3. Perform exact inference only using the pool of candidates.
4. Perform m local optimizations of a single pitch track given the other ones.

The algorithm has complexity of $O(mn_\omega^2 T)$ for the Viterbi algorithm with single pitch tracks, and $O(Tp_\omega^m)$ for the structured Viterbi algorithm of step 2. In practise, p_ω is small enough (around 10) so that step 3 is not the bottleneck while being large enough to yield no significant difference from the setting $p_\omega = n_\omega$ (i.e., no approximation).

4. LEARNING OF PARAMETERS

If we denote $z = (\omega, v, h, c, \lambda)$, then we have a model for s which is a latent variable model with latent variable z. In the presence of "labelled data," i.e., datasets for which both s and z are available, there are two different types of training that can be employed, generative or discriminative.

In this paper, we use pitch-labelled data from the Keele pitch database [1]. This database has ten different speakers; the pitch frequency ω and the voicing decision v are available, but neither the harmonic amplitudes h nor the unvoiced constant amplitude c are available

We can create artificial labelled training data with several speakers by superposing two distinct signals. In this paper, we consider mixing of two speakers for training and mixing of two or three speakers for testing (note that since the parameters are shared by all speakers in our framework, learning only on two speakers leads to a pitch extractor that can deal with any number of pitches). We thus have two sets of hidden variables (ω^1, v^1), (ω^2, v^2), one for each speaker.

4.1. Generative training (maximum likelihood)

In this type of estimation, if we have observations for both x and the hidden states z, we simply maximize the joint likelihood $p(s, z)$ of the data (s, z). Since we have a directed graphical model, this readily decouples in independent parameter estimations for each conditional distribution [9]. The data from Keele pitch database do not include the harmonic amplitudes; the harmonic amplitudes that do not correspond to the pitch value ω (which is observed) do not play any role in the model, thus they can be left unspecified; for the harmonic amplitudes corresponding to the observed pitch, we take h to be the best amplitudes in the least-square sense, i.e., $h_\omega = B(\omega)^\top (B(\omega)^\top B(\omega))^{-1} s$.

Although efficient (no inference in an HMM has to be performed for learning), such training, when the final objective of inference is only to estimate the hidden state z and not to also obtain a model of the observations, is usually outperformed by discriminative training, which directly optimizes the conditional likelihood $p(z|s)$ [14, 10].

4.2. Discriminative training

Instead of maximizing $p(s, z)$, we maximize the conditional likelihood $p(z|s)$. Maximizing the conditional likelihood does not decouple in a graphical model and thus exact maximum conditional likelihood estimation requires performing many runs of the inference algorithm for factorial HMMs, even to simply compute $p(z|s)$. Since exact inference is intractable in those HMMs, we instead maximize a "pseudo log likelihood" which is defined as the sum of the log likelihoods of subproblems and exhibits asymptotic properties similar to full maximum likelihood [15]. We defined the pseudo log likelihood as follows: the available data is $(\omega^1, v^1, \omega^2, v^2)$; we let $q(\omega, \omega', v, v')$ denote

$$q(\omega, \omega', v, v') = \max_{h^1, h^2, c^1, c^2, \lambda} p(\omega, \omega', v, v', h^1, h^2, c^1, c^2, \lambda).$$

We maximize with respect to the parameters the log likelihood defined as:

$$\log \frac{q(\omega^1, \omega^2, v^1, v^2)}{\sum_{\omega, v} q(\omega, \omega^2, v, v^2)} + \log \frac{q(\omega^1, \omega^2, v^1, v^2)}{\sum_{\omega, v} q(\omega^1, \omega, v^1, v)}$$

The maximization is performed through gradient descent, and requires inference in an HMM with a number of states proportional to n_ω, as opposed to n_ω^2.

5. SIMULATIONS

In this section, we show that the various features that were included into our graphical model framework lead to robust performance. In all our simulations, training was performed on the first 6 speakers, while testing was performed on the remaining 4 speakers. The metric we use to compare pitch frequencies ω and ω' is $d(\omega, \omega') = 1 - e^{-(\omega - \omega')^2/\sigma^2}$, where σ^2 is the empirical variance of the pitch frequency over the entire training set. This measure is equivalent to the squared distance for close values and tends to 1 for distant values. We prefer it to the plain squared distance, because if an estimated pitch is far away from the true pitch, its value is not relevant and we prefer to have a fixed unit penalty for all clearly wrong values of the pitch.

The running time for extracting any number of pitches is linear in the duration of the signals. In our current Matlab implementation with a 2GHz processor, the running time is 30 times the duration of the signal for extracting one pitch, while it is 130 times the duration of the signal for extracting two pitches.

5.1. Effect of smoothing spline prior

For the simple task of pitch determination for independent frames taken from one speaker, we have compared our approach to an

	voicing	pitch error
female - male	22%	0.03
female - female	32%	0.08
male - male	31%	0.07

Fig. 4. Double pitch extraction: voicing decision error rates and mean pitch estimation errors.

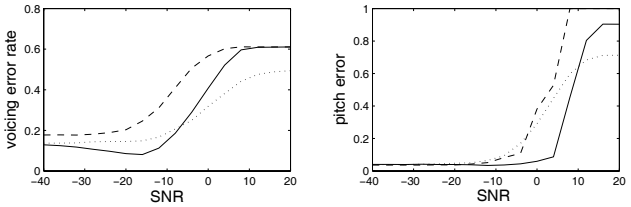

Fig. 5. Single pitch extraction with noise: voicing decision error rates (left) and mean pitch estimation errors (right); white noise (plain), stationary colored noise (dashed), restaurant background (dotted).

approach without a spline smoothing prior on the harmonic amplitudes: with the smoothing spline prior, the average error on the pitch estimate using the measure defined earlier is equal to 0.28, while the error for the estimate without smoothing spline prior is equal to 0.57, and most of the additional errors are due to pitch halving, which is a well known problem in pitch determination. In the context of harmonic modeling approaches such as the one presented in this paper, a priori detailed knowledge of the spectral envelope has been shown to remove the pitch halving ambiguity [3]; the current results suggest that a simple spline smoothing prior which does not require knowledge of the envelope is also sufficient to resolve this ambiguity.

5.2. Discriminative vs. generative training

On single pitch tracking experiments, we compared the performance of pitch extractors trained discriminatively or generatively. The pitch extractor trained generatively made an incorrect decision regarding voicing 27.4% of the time and had a pitch estimation error of 0.022, while the pitch extractor trained discriminatively made an incorrect decision regarding voicing only 5% of the time and had a pitch estimation error of 0.016. Discriminative training indeed leads to significantly better performance.

5.3. Two speakers

In this set of experiments, we mixed two signals from different speakers with same energy. In Figure 4 we report incorrect voicing decision rates and mean pitch estimation errors, with speakers of different genders.

5.4. Noisy conditions

We also performed experiments in which we added three different types of noise to the signal: white noise, stationary colored noise and non-stationary restaurant background noise. We plot the results as a function of signal-to-noise ratios in Figure 5, illustrating the robustness to noise of our pitch extractor.

6. CONCLUSION

We have presented an algorithm for multiple pitch extraction based on graphical models. The use of appropriate prior distributions and discriminative training leads to robust extraction performance. Importantly, the computational complexity of our algorithm is linear in the length of the audio segment. Although the running time of our current Matlab implementation is 130 times slower than real time, we do not foresee any major obstacles to the design of a more efficient software implementation that runs in real time.

7. REFERENCES

[1] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. EUROSPEECH*, 1995.

[2] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, Wiley Press, 1999.

[3] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. ICASSP*, 1990.

[4] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 3, 2003.

[5] X. Li, J. Malkin, and J. Bilmes, "Graphical model approach to pitch tracking," in *Intl. Conf. Spoken Lang. Proc.*, 2004.

[6] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters," in *Proc. IEEE Work. App. Sig. Proc. Acoust.*, 1999.

[7] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Proc.*, vol. 12, no. 1, 2004.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[9] M. I. Jordan, "Graphical models," *Stat. Sci.*, vol. 19, no. 1, pp. 140–155, 2004.

[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001.

[11] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[12] G. Wahba, *Spline Models for Observational Data*, SIAM, 1990.

[13] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[14] L. R. Bahl, P. V. de Souza P. F. Brown and, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, 1986.

[15] G. Liang and B. Yu, "Maximum pseudo likelihood estimation in network tomography," *IEEE Trans. Sig. Proc.*, vol. 51, no. 8, pp. 2043–2053, 2003.