PITCH ESTIMATION USING MODELS OF VOICED SPEECH ON THREE LEVELS

Dominik Joho, Maren Bennewitz, and Sven Behnke

University of Freiburg, Germany
Department of Computer Science
{joho, maren, behnke}@informatik.uni-freiburg.de

ABSTRACT

We present an algorithm for estimating the fundamental frequency in speech signals. Our approach incorporates models of voiced speech on three levels. First, we estimate the pitch for each time frame based on its harmonic structure using non-negative matrix factorization. The second level utilizes temporal pitch continuity to extract partial pitch contours. Thirdly, we incorporate statistics of the succession of voiced segments to aggregate partial contours to the final contour of an utterance. We evaluate our approach on the Keele database. The experimental results show the robustness of our method for noisy speech, and the good performance for clean speech in comparison with state-of-the-art algorithms.

Index Terms— pitch estimation, speech analysis, matrix decomposition

1. INTRODUCTION

One of the most salient features of human speech is its harmonic structure. During voiced speech segments, the regular glottal excitation of the vocal tract produces energy at the fundamental frequency (F_0) and its multiples. The changing pitch carries a good part of the auditory message. It discriminates words in tonal languages, allows expressing emotions, discriminates questions from statements, and allows emphasizing parts of an utterance. Furthermore, pitch tracking is the basis for the separation of harmonic speech from other speech components and background noise [1].

In this paper, we present and evaluate an algorithm for estimating the fundamental frequency or pitch in speech signals. Our algorithm utilizes models of voiced speech on three levels. The first level is the time frame for spectral analysis. We decompose the short term spectrum using non-negative matrix factorization (NMF). The spectrum is represented as a weighted sum of harmonic templates and templates for non-harmonic speech. NMF outputs a matrix that holds harmonic energy along possible pitch contours. Then a hidden Markov model (HMM) extracts a set of partial contours from this matrix. This step makes use of the time continuity of F_0 in voiced segments. On the third level, we rely on statistics of the succession of voiced segments to combine a subset of the partial contours to the final pitch contour of an utterance. The subset is determined by policy iteration.

As NMF decomposes a spectrogram into additive parts, it is usable for auditory scene analysis [2], where multiple auditory objects have to be separated. In principle, NMF with the proposed templates can also be applied to multi-pitch tracking and simultaneous tracking of formants. In this paper, however, we concentrate on tracking a single speaker reliably.

We present experiments on the Keele database showing the robustness of our method under several acoustic conditions and the competitiveness with state-of-the-art algorithms.

The next section gives an overview of related work and Section 3 introduces NMF. Section 4 describes the proposed algorithm in detail. In Section 5 we present evaluation results and compare them with results of existing algorithms.

2. RELATED WORK

The problem of pitch estimation has been addressed for a long time using many different approaches. In recent years, techniques like statistical learning [3, 4], time domain probabilistic approaches for waveform analysis [5], or optimization techniques [6] have been applied to accomplish this task. However, most of these techniques are not robust enough, especially for corrupted speech.

NMF has been used for various problem domains, such as face detection and semantic analysis of text documents [7], polyphonic music transcription [8], as well as discovery of hierarchical speech features [9]. Sha and Saul [3] used NMF for pitch estimation by utilizing an instantaneous frequency based representation of the speech signal. They used basis vectors for fundamental frequencies ranging from 50 Hz up to 400 Hz and one non-harmonic basis vector. Their approach focused on statistical learning of the basis vectors by using reference data. Our approach is motivated by the source filter model of speech, uses the log-spectrogram as speech representation, and a different set of given basis vectors. Furthermore, we implemented a second and third level of analysis, in order to be able to extract pitch contours even in case of overlapping speakers.

This work has been supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), grant BE 2556/2-1.

1-4244-0728-1/07/$20.00 ©2007 IEEE

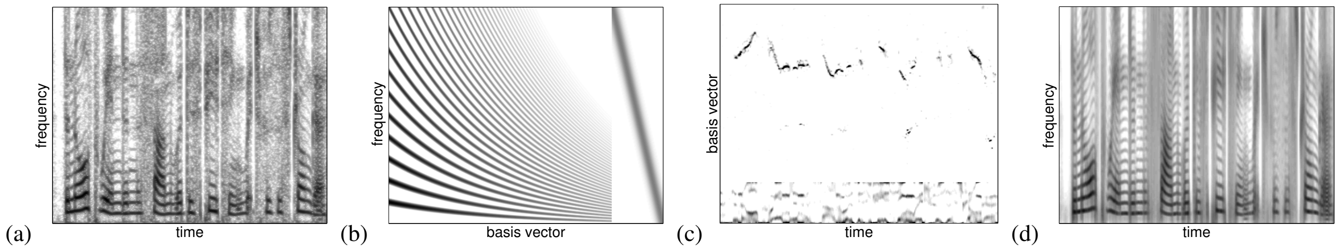

Fig. 1. Non-negative matrix factorization: (a) The V matrix is the log-spectrogram of a speech signal. (b) The W matrix with basis vectors for different fundamental frequencies (left part) and for modeling the frequency response of the vocal tract filter (right part). (c) The resulting H matrix contains preliminary information about the pitch contour (upper part) and formants (lower part). (d) The resulting approximation $\widehat{V} = WH \approx V$ looks like a smoothed version of the original spectrogram (a).

3. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization [7] decomposes a given non-negative matrix $V^{n\times m}$ into two non-negative matrices $W^{n\times k}$ and $H^{k\times m}$, such that $V \approx WH$. The m columns of V consist of n-dimensional data vectors. The k columns of W contain basis vectors of dimension n. Each n-dimensional column vector of the approximation $\widehat{V} = WH$ is a linear combination of all basis vectors, whereby the coefficients are the entries of the corresponding k-dimensional column vector of H. One measure of the factorization quality is the divergence, which is defined as:

$$D(A||B) = \sum_{ij}(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}). \qquad (1)$$

$D(V||WH)$ is minimized by

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \qquad (2)$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} V_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}}. \qquad (3)$$

Lee and Seung [10] proved that these update rules find local minima of their objective function $D(V||WH)$. The multiplicative update does not change the sign of W or H. Hence, if they are initialized to positive values, no further constraints are necessary to enforce their non-negativity.

4. PITCH ESTIMATION

We compute the spectrogram of the signal by using a window length of 51.2 ms and a shift of 10 ms. Frequencies above 4 kHz are cut off, because most of the information usable for estimating pitch is contained in frequency-bands below 4 kHz. Reducing the size of the spectrogram also improves the run-time of the algorithm. Our method to estimate pitch contours consists of the following three steps.

4.1. Factorizing the Spectrogram

The spectrogram is interpreted as V matrix and will be factorized by the NMF. The W matrix is given and fixed, so that only the update rule (2) for H is applied. The basis vectors of W describe the additive parts we expect to find in V. They can be divided in two categories: Harmonic basis vectors, which model the spectrum of the excitation signal for voiced speech at different fundamental frequencies, and filter basis vectors, which are used to model the frequency response of the vocal tract filter (see Fig. 1(b)). The harmonic basis vectors consist of several scaled log-normal distributions (with varying variance and scale factors) centered around their respective fundamental frequency and their integer multiples. We used 278 basis vectors for modeling fundamental frequencies ranging on a log-scale from 50 Hz to 400 Hz. The filter basis vectors were modeled by using 64 basis vectors, each containing a binomial distribution centered around different frequencies and one uniform basis vector for plosives. Ideally, each time-frame of voiced speech would be approximated by one harmonic basis vector and several filter basis vectors, while the approximation for unvoiced time-frames will lack the use of a harmonic basis vector.

According to the source filter model, the spectrum of the excitation signal and the frequency response of the vocal tract filter are combined multiplicatively, but the matrix multiplication WH combines the basis vectors in a linear combination. Hence, the logarithm of both, the spectrogram and the basis vectors has to be taken in order to actually turn the addition of the linear combination into a multiplication. After applying the iterative update rule (2) to estimate H, the upper part H_f of H contains preliminary information about the pitch contour (see Fig. 1(c)). By using the information in the lower part of H, it might be possible to track the formants. However, we did not investigate in that direction.

4.2. Extracting Voiced Segments

We use an HMM [11] to track pitch contours as Viterbi paths in H_f. This is motivated by the continuity of F_0 within voiced segments. A Viterbi path is the most likely sequence of hidden states, which in our case correspond to the fundamental frequency. Instead of computing one pitch contour along the full time axis of the matrix, several short pitch contours are extracted, which will be called partial contours. These contours represent voiced segments within H_f. The maximum

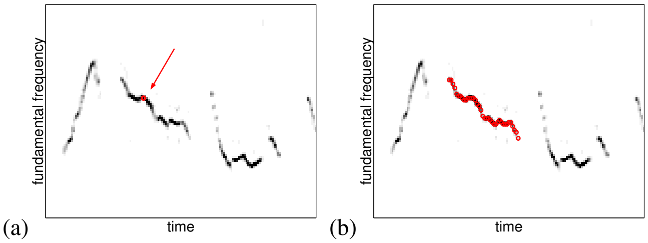

Fig. 2. (a) Finding the maximum of H_f as starting point for extraction of a partial contour. (b) Determining Viterbi paths leaving the maximum backwards and forwards in time.

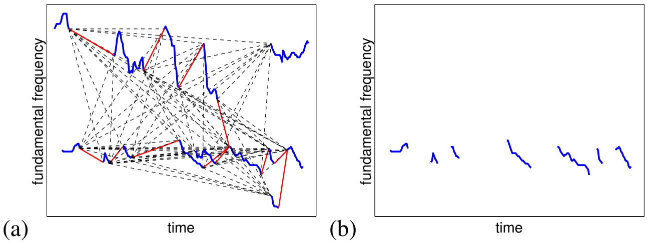

Fig. 3. (a) Partial contours indicating possible successors (black, dashed) and forward policy (red, solid). (b) Resulting pitch contour if the start point would be the lower, leftmost partial contour.

of H_f is determined (see Fig. 2(a)) as a starting point for the computation of two Viterbi paths, one leading forward in time and one backwards (see Fig. 2(b)). As a prior, we set the probability of being in the state of the maximum to one and zero for all other states. For transition and observation probabilities we use binomial distributions that we estimated from speech data.

The computation of the Viterbi path in each direction terminates if the most likely state s_c at the current point in time t_c is likely to be unvoiced. Let (s_0, t_0) denote the position of the maximum and $\theta > 0$ be a threshold. Then a voicing criterion can be specified as

$$\frac{H_f(s_c, t_c) - mean(H_f)}{H_f(s_0, t_0) - mean(H_f)} < \theta \qquad (4)$$

The maximum and the two outgoing Viterbi paths then constitute a partial pitch contour, and the next partial contour can be determined. In order not to find the same maximum again, the partial contour has to be set to zero in H_f. Ideally, it would be sufficient to set all matrix entries of this partial contour to zero. However, the footprint of a pitch contour in H_f is somewhat blurred along the pitch axis and the partial contour also has to be set to zero along the pitch axis, using the same voicing criterion.

4.3. Selecting a Subset of Partial Contours

The resulting set $\mathcal{C}$ of detected partial contours may still contain false positives and – for overlapping speakers – partial contours of different speakers. In order to find a reasonable subset $\widehat{\mathcal{C}} \subseteq \mathcal{C}$ that corresponds to the pitch contour of the dominant speaker, we estimate his or her global pitch first. A weighted histogram is built, where each partial contour votes for its average pitch frequency and weights its vote with its average energy (average of its entries in H_f). The histogram is then smoothed. Partial contours that are false positives or do not belong to the dominant speaker will have a lower energy in average. Hence, the global maximum of the histogram is defined as the global pitch f_g of the dominant speaker.

The decision of which partial contour should be contained in the subset $\widehat{C}$ can be made according to two types of criteria: The ones that are based solely on a partial contour itself, e.g., its average weight, and criteria that aim to assure a reasonable course of the pitch by restricting the possible successors of a partial contour. These criteria can be incorporated in the idea of "walking" through the partial contour. The best path will be determined by using policy iteration (PI) [11]. Partial contours are states s, and actions $a_{s'}$ are "jumps" to a partial contour s' in the neighborhood of s. A reward function

$$R(s) = w_a(s) \cdot w_t(s) \cdot (w_a(s) \cdot \mathcal{N}(f_c(s), f_g, \sigma_{f_g})) \qquad (5)$$

specifies how desirable it is to have partial contour s as part of the final pitch contour. It is defined using the contour's average weight $w_a(s)$ and total weight $w_t(s)$. The frequency $f_c(s)$ of its center point is used to weight the reward by a normal distribution $\mathcal{N}(x, \mu, \sigma)$ with experimentally determined variance σ_{f_g} and mean at the global pitch f_g in order to favor partial contours nearby the global pitch. The transition probability function

$$T(s'|s, a_{s'}) = w_a(s') \cdot \Delta(\delta_f, \delta_t) \qquad (6)$$

specifies how likely one considers a jump from partial contour s to partial contour s', and is defined by the average weight $w_a(s')$ of s' and a jump probability $\Delta(\cdot, \cdot)$[1]. As each jump covers distances δ_t along the time axis and δ_f along the fundamental frequency axis, the jump probability is defined as a two-dimensional normal distribution using pitch contours of reference data.

Two optimal policies are determined by policy iteration, one for walking forward in time, and one for walking backwards (see Fig. 3(a)). The partial contour with the highest $R(s)$ is defined to be the start point. Its predecessors and successors are given by walking through the partial contours using the two policies. Each traversed partial contour is defined to be part of $\widehat{C}$. In this way, the pitch contour is estimated (see Fig. 3(b)).

5. EXPERIMENTS

To evaluate the performance of our algorithm, we used the Keele pitch reference database [12]. This database consists of speech signals of five male and five female English speakers each reading the same phonetically balanced text with varying duration between about 30 and 40 seconds. The reference

[1] $T(s''|s, a_{s'})$ is zero for all states $s'' \neq s'$, except for a global zero-reward terminal state s_0: $T(s_0|s, a_{s'}) = 1 - T(s'|s, a_{s'})$. This state does not have to be modeled explicitly.

Table 1. Evaluation results for different types of noise at varying signal-to-noise ratios.

	white noise					cocktail party noise					second speaker			
	VE	UE	GPE	RMS		VE	UE	GPE	RMS		VE	UE	GPE	RMS
20 dB	7.6	2.5	0.9	3.8	20 dB	7.8	5.4	1.5	4.0	20 dB	8.7	7.2	2.3	3.9
15 dB	9.1	2.5	1.1	3.8	15 dB	9.1	7.0	1.6	4.5	15 dB	8.5	8.1	3.5	4.3
10 dB	11.6	3.0	1.0	3.9	10 dB	12.1	7.7	2.5	5.1	10 dB	9.8	8.8	5.2	4.5
5 dB	15.4	3.8	1.2	4.5	5 dB	18.2	9.4	3.6	6.2	5 dB	12.7	11.1	17.0	5.2
1 dB	24.6	3.1	1.1	4.8	1 dB	28.8	7.4	3.2	7.1	1 dB	15.5	13.7	32.7	6.2

pitch estimation is based on a simultaneously recorded signal of a laryngograph. Uncertain frames are labeled using a negative flag. The authors of the database suggest to ignore these frames in performance comparisons.

We use common performance measures for comparing pitch estimation algorithms: The voiced error (VE) denotes the percentage of voiced time frames misclassified as unvoiced, the unvoiced error (UE) is defined as the inverse case, the gross pitch error (GPE) denotes the percentage of frames at which the estimation and the reference pitch differ by more than 20%, and the root mean square error (RMS) is computed as RMS difference in Hertz of the reference pitch and the estimation for all frames that are not GPEs.

Tab. 2 presents evaluation results of the proposed algorithm (NMF-HMM-PI) for clean speech. For comparison we list results of other state-of-the-art algorithms [3, 4, 5, 13] that are based on the same reference database. As can be seen, our method yields very good results in comparison with these approaches.

To test the robustness of the algorithm, we added different types of noise at different signal-to-noise ratios (SNRs) to the original signals. Noise types we used were white noise, cocktail party noise, and a second speaker of the opposite sex. As can be seen in Tab. 1, difficult SNRs seem to have a stronger impact on the VEs than on the GPEs for white noise and cocktail party noise. The high GPEs for mixed signals with a second speaker at 5 db and 1 dB are due to tracking the other speaker in some of the test signals.

6. CONCLUSION

In this paper, we presented a pitch estimation method that relies on three levels of analysis. The first level separates the periodic excitation signal and the frequency response of the vocal tract filter by using NMF in a physically plausible manner. The second level uses an HMM to aggregate harmonic energy in voiced segments to partial pitch contours. Finally, we use PI to incorporate statistics of the succession of voiced segments to select a subset of the partial contours for the final pitch estimate of an utterance. We evaluated our algorithm for clean speech as well as for demanding acoustic conditions. The experimental results show the competitiveness with state-of-the-art methods for clean speech and the robustness under difficult acoustic conditions.

Table 2. Evaluation results of our algorithm (NMF-HMM-PI) for clean speech. For comparison we list results found in the literature.

	VE	UE	VE+UE	GPE	RMS
NMF-HMM-PI	7.08	**2.43**	**9.51**	1.06	**3.66**
NMF [3]	7.7	4.6	12.3	**0.9**	4.3
RAPT [3]	**3.2**	6.8	10.0	2.2	4.4
MLS$^+$ [4]	7.03	7.90	14.93	1.50	4.54
Seg. HMM [5]	8.49	8.89	17.38	2.28	4.48

7. REFERENCES

[1] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, 2001.

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

[3] F. Sha and L. K. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS) 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds., pp. 1233–1240. MIT Press, 2005.

[4] F. Sha, J. A. Burgoyne, and L. K. Saul, "Multiband statistical learning for F0 estimation in speech," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, 2004, pp. 661–664.

[5] K. Achan, S. Roweis, A. Hertzmann, and B. Frey, "A segment-based probabilistic generative model of speech," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 221–224.

[6] L. K. Saul, D. D. Lee, C. L. Isbell, and Y. LeCun, "Real time voice processing with audiovisual feedback," in *Advances in Neural Information Processing Systems (NIPS) 15*, S. Becker, S. Thrun, and K. Obermayer, Eds., pp. 1181–1188. MIT Press, 2003.

[7] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[8] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Appl. of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[9] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2003, vol. 4, pp. 2758–2763.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS) 13*, 2001, pp. 556–562.

[11] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2. edition, 2003.

[12] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. of Eurospeech*, 1995, vol. 1, pp. 837–840.

[13] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Eds. Elsevier, 1995.