DETECTING PITCH OF SINGING VOICE IN POLYPHONIC AUDIO

Yipeng Li and DeLiang Wang

Department of Computer Science and Engineering
& Center of Cognitive Science
The Ohio State University
Columbus, OH, 43210-1277,USA
{liyip,dwang}@cse.ohio-state.edu

ABSTRACT

We propose a robust algorithm to detect the pitch of singing voice in polyphonic audio. A new channel/peak selection scheme is introduced to exploit the salience of singing voice and the beating phenomenon in high frequency channels. An HMM model is employed to integrate the periodicity information across frequency channels and time frames. Quantitative evaluation shows that the new system performs significantly better than existing algorithms for predominant pitch detection in polyphonic audio.

1. INTRODUCTION

Although monaural sound segregation remains a great computational challenge, the human auditory system shows a remarkable capability in this task. One well known example is that a listener can separate speech from the background noise in a cocktail party. Another example is that a listener can hear out the singing voice that is accompanied by musical instruments. The goal of the work presented here is to mimic this functionality of the human auditory system in separating singing voice from the accompaniment in real polyphonic audio. Our approach is to first find the pitch contour of the singing voice and then use the detected pitch contour to segregate the singing voice from the accompaniment. This approach is motivated by psychoacoustic evidence that pitch is crucial in the perception and organization of sound. A recent system of monaural speech segregation by Hu and Wang [1] demonstrated the effectiveness of pitch-based segregation for voiced speech. This paper focuses on detecting the pitch of singing voice, which is called predominant pitch detection.

The detection of the pitch of a harmonic sound in the presence of other sounds has been mainly studied in the context of speech segregation, speech enhancement and other related fields. Before applying techniques developed for predominant pitch detection for speech, it is instructive to make a comparison between singing and speech.

As two main forms of human voice, singing and speech are similar in many aspects. For example, they are both harmonic for voiced sounds and composed of both voiced and unvoiced sounds. But the differences between singing, especially operatic singing, and speech are also significant. A well known difference is the presence of an additional formant, called the singing formant, at frequencies between 2000–3000 Hz. This singing formant helps the voice of a singer stand out of the accompaniment [2]. Since the singing formant makes certain frequency components of the singing voice more salient, it can be used to facilitate predominant pitch detection. Another difference is in the dynamics of pitch. During singing, a singer usually intentionally stretches the voiced sounds and shrinks the unvoiced sounds to match sounds from other musical instruments. As a result, the pitch contour of the singing voice tends to be relatively piece-wise constant. This difference could be important in adapting predominant pitch detection algorithms, which use the pitch dynamics (continuity constraint) of speech to guide pitch detection or reduce the error. Besides these two differences, singing also has a wider pitch range. The pitch range of normal speech is between 80 and 400 Hz while that of singing is between 80 and 1000 Hz. Better frequency resolution is required to detect high pitch.

Other difficulties in detecting the pitch of singing voice arise from the presence of accompaniment in polyphonic audio. One complication comes from the broadband nature of the accompaniment, which makes the frequency components of the voice and music overlap significantly. Another complication is that multiple melodies might exist concurrently, the singing voice being one of them, in polyphonic audio. Identifying the melody carried by singing voice can be difficult.

Recently several predominant pitch detection algorithms have been proposed for polyphonic audio. Shandilya et al. [3] used the pitch perception model of Meddis and Hewitt [4] to detect the pitch of singing voice in the presence of percussive sounds for Indian film songs. The frequency corresponding to the lag of the highest peak in the summary autocorrelation function is considered the predominant pitch. Goto [5] used the Expectation-Maximization algorithm to estimate the probability of each frequency being the funda-

mental from a filtered spectrum. The most likely frequency is considered as the pitch of the singing voice. His system was tested on real commercial recordings and 80% accuracy was reported for melody line and bass line detections.

Multipitch detection algorithms can also be used for predominant pitch detection. The first pitch being detected is considered the predominant pitch. For example, the multipitch detection algorithm proposed by Klapuri [6] detects the pitch of each instrument in the mixture of steady musical sounds using the principles of harmonicity and spectral smoothness. The first detected pitch is predominant in the sense that the score of this pitch hypothesis is the highest.

In this paper, we propose an algorithm for detecting the pitch of singing voice in polyphonic audio. Our algorithm extends the one by Wu et al. [7], which was designed for multipitch detection for noisy speech. The input signal is first filtered by an auditory periphery and a correlogram is calculated to extract the periodicity information in each channel. Channel and peak selection is then applied to obtain useful periodicity information, which is integrated by a statistical model. Finally an HMM is used to model the pitch generation process and the most probable pitch track is identified as the pitch contour of singing voice. Frequency decomposition and the simultaneous detection of two pitch contours help to reduce the interference of some other melody from the accompaniment. The HMM framework makes the algorithm robust in the presence of strong musical interference and easy to incorporate the pitch dynamics of singing voice.

The paper is organized as the following. Section 2 gives a detailed description of our system. Evaluation is given in Section 3. The last section concludes the paper.

2. SYSTEM DESCRIPTION

Our algorithm consists of five stages. The first stage is the auditory periphery [8]. The signal is sampled at 16 kHz and passed through a 128-channel gammatone filterbank. The channels are classified into low frequency channels (center frequency below 800 Hz) and high frequency channels (center frequency above 800 Hz). In each high frequency channel, the envelope of the output is extracted.

After the auditory filtering, a normalized autocorrelation is computed in the second stage for each channel with a frame length of 16 ms and the frame shift of 10 ms to obtain the periodicity information. For low frequency channels the autocorrelation is computed directly on the output of a filter while this is done on the output envelope for high frequency channels. The peaks in the correlogram contain the periodicity information to be used in a later stage for pitch hypotheses evaluation. However, the existence of other concurrent sound sources makes the peaks of the autocorrelation in some channels misleading.

In the third stage, we apply channel selection in the low frequency range to identify clean channels. A low frequency channel will be selected if the maximum value of the autocorrelation in the plausible pitch range exceeds a threshold of $\theta = 0.945$. In the high frequency range, we retain all the channels and apply peak selection to make use of the beating phenomenon. Specifically, only the first peak of the autocorrelation is retained for each channel.

Not applying channel selection makes more high frequency channels available, which is important in distinguishing different harmonic sources. However, it also introduces noisy peaks, whose lags do not correspond well to the fundamental period of the singing voice. But we have found experimentally that the lag of the first peak, within the pitch range in a noisy high frequency channel, is still a good indicator of the true pitch of singing voice in many cases. This is not caused by the singing formant since for the genres we tested the singing formant is not present. However it might be the case that the high frequency components become more salient because of singing. It is also well known that high-frequency channels respond to multiple harmonics and the envelope of the response fluctuates at the fundamental frequency [9] . We therefore discard the channel selection and select only the first peak of the autocorrelation in high frequency channels. Compared to the system in [7] this makes available the high-frequency information which is important as stated previously. As a result, this technique along with the following statistical cross-channel integration greatly improves the performance of detecting pitch of singing voice (see Section 3).

The statistical integration across frequency channels and time frames is formulated in a similar way to [7]. In the fourth stage of our algorithm, the score of a pitch hypothesis is calculated. Notice that, if a channel is clean, the distance, δ, between the lag of the true pitch and that of the closest observed peak tends to be small. This relation can be quantitatively described by a Laplacian distribution, which centers on zero and exponentially decreases as the absolute distance of the two lags increases:

$$L(\delta;\lambda_c) = \frac{1}{2\lambda_c}\exp\left(-\frac{|\delta|}{\lambda_c}\right) \quad (1)$$

where the distribution parameter λ_c is a function of the channel center frequency (related to channel number).

Channels not selected ("background noise" channels) are modelled by a uniform distribution $U(\delta;\eta_c)$ where η_c indicating the possible range of the distance of the two lags. The observation probability of 1-pitch hypothesis for channel c is given by

$$p_c\left(\Phi_c|d\right) = p_c\left(\delta\right) = \left(1-q\right)L\left(\delta;\lambda_c\right) + qU\left(\delta;\lambda_c\right) \quad (2)$$

where Φ_c is the set of selected peaks in channel c and d is the lag of the true pitch. q is the partition factor $(0 < q < 1)$.

The observation probability of 2-pitch hypothesis can be

formulated based on that of 1-pitch:

$$p_c\left(\Phi_c|\left(d_1,d_2\right)\right)=\begin{cases} q_2\left(c\right)U\left(0;\eta_c\right) \\ \quad \text{if channel } c \text{ is not selected} \\ p_c\left(\Phi_c|d_1\right) \\ \quad \text{if channel } c \text{ belongs to } d_1 \\ max\left(p_c\left(\Phi_c|d_1\right),p_c\left(\Phi_c|d_2\right)\right) \\ \quad \text{else} \end{cases} \tag{3}$$

where $q_2(c)$ is the partition factor for channel c under 2-pitch hypothesis. Channel c belongs to the source d_1 if the distance between d_1 and the closest peak in channel c is less than $5\lambda_c$. All parameters are obtained from clean signals using the maximum likelihood method in a manner similar to [7].

The score for a given pitch hypothesis across all channels is given by (using a 2-pitch as an example):

$$p\left(\Phi|\left(d_1,d_2\right)\right)=k_2\sqrt[b]{\prod_{c=1}^{C}p_c\left(\Phi_c|\left(d_1,d_2\right)\right)} \tag{4}$$

where Φ is the set of all selected peaks. C is the total number of channels and k_2 is the normalization factor. b is used to compensate for statistical dependency among channels.

The final stage of the system performs pitch tracking by an HMM, which approximates the pitch generation process. We define the pitch state space as the union of three i-dimensional subspaces ($i = 0, 1, 2$), each of which represents the collection of hypotheses with i pitches. In each frame, a hidden node represents the pitch state space and the observation node represents the set of observed peaks. The transition between consecutive frames, i.e., between different states in pitch space, is described by pitch dynamics, which is composed of two parts: the transition probability between different pitch subspaces, and that between different pitch configurations in the same pitch subspace. The transition behavior within the same pitch subspace can also be described by a Laplacian distribution, while the transition probability between different subspaces can be determined by training. Finally, the Viterbi algorithm is used to find the most likely sequence of pitch generation and transition. The first track detected is considered as the predominant pitch contour corresponding to singing voice.

3. RESULTS AND COMPARISON

The test samples are extracted from commercial CD recordings. Some modern karaoke CDs are recorded with multiplex technology. The singing and the accompaniment are multiplexed and stored in the same channel. With proper demultiplexing software, the separate singing voice can be extracted for the evaluation of pitch detection as well as sound separation. In order to demonstrate the applicability of the proposed algorithm on a wide range of polyphonic audio, we extract a total of 25 clips from 9 songs belonging to three genres: country, pop and rock. The total length of the clips is about 35 seconds. The singing voice and the accompaniment are mixed with an overall SNR 0 dB. In this case, the music is strong while listeners can still hear the singing voice clearly.

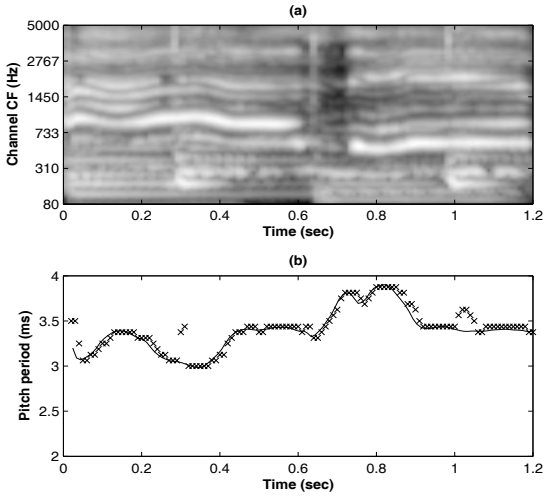

Fig. 1. Predominant pitch detection on a clip of country music. (a) Time-frequency energy plot of the clip. Brighter area indicates stronger energy. The vertical axis shows the central frequencies of channels. (b) Result of pitch detection. The solid line indicates the reference pitch and "×" represents the detected pitch.

The reference pitches are calculated using Praat [10]. The clean singing voice is processed by Praat and the resulted pitch track is visually inspected for obvious pitch halving and doubling errors.

Fig. 1 shows the result of pitch detection on a clip of country music. The energy plot of the clip is shown in Fig. $1(a)$. In this example, the singing voice is dominant in high frequency channels while the low frequency channels are severely corrupted by accompaniment. The detected pitch is plotted against the reference pitch in Fig. $1(b)$. In this example, the detected pitch track well matches the reference track.

The overall performance of the proposed system on the testing database is shown in the last column of Table 1 in terms of gross error rate. A gross error occurs if the detected pitch is not within 20% of the reference pitch. In the table, the error rates are grouped by genres and the average gross error rate for all three genres is listed.

We have compared the performance of our algorithm with that of the following predominant pitch detection algorithms. A correlogram method similar to that in [4] is used to compute a summary autocorrelation and the lag of the most salient peak in the summary autocorrelation, within

Table 1. Gross error rate for different methods

GENRE	METHOD			
	Correlogram	Klapuri	Wu, et al.	Proposed
country	68.1	41.0	35.1	16.2
pop	64.9	35.5	18.0	17.6
rock	86.7	57.2	72.6	14.7
	73.6	45.3	44.3	15.9

the plausible pitch range, provides the estimated pitch of the singing voice. Klapuri's multipitch detection algorithm is specifically designed for music. It can detect up to 6 concurrent steady musical sounds and performs relatively well in the presence of percussive sounds. We have implemented this algorithm and adapted it to detect the pitch of singing voice using a 20 ms window. The window size used in [6] is 190 ms, which is unsuitable for the non-stationary polyphonic audio studied in this paper. The first pitch detected by this algorithm is used as the predominant pitch. We also show the performance of the system in [7] to illustrate the improvement made by the proposed system. For comparison purposes, the pitch range for all algorithms is set to 80–500 Hz. The comparative results are given in Table 1.

The correlogram method performs poorly in predominant pitch detection of all the three categories when multiple sounds are present. This shows that the most salient peak in the summary autocorrelation function is not a reliable indicator for singing voice in polyphonic audio. The performance of Klapuri's algorithm is better compared to that of the correlogram method but still unsatisfactory. The algorithm by Wu et al. performs well for pop music, but its performance on rock music is particularly poor because of the under-utilization of the periodicity information in high frequency channels. In the presence of strong percussive sounds as encountered in country and rock music, the low frequency channels do not provide enough information in distinguishing different sound sources. In this case, the periodicity information in high frequency channels becomes crucial because the singing voice usually dominates in those channels. The successful recovery of such information gives our algorithm a superior performance in country and rock music and significantly better overall pitch detection accuracy.

4. CONCLUSION

In this paper we have proposed a predominant pitch detection algorithm for detecting the pitch of singing voice in monaural polyphonic music. The new channel/peak selection method introduced here utilizes the salience of the singing voice and the beating phenomenon in high frequency channels. This as well as the statistical integration of periodicity information using HMM greatly improves the accuracy of predominant pitch detection for singing voice. The performance of the proposed algorithm is significantly better than other predominant pitch detection algorithms tested. The detected pitch of singing voice can be readily fed to a pitch-based sound separation system to segregate the singing voice from the accompaniment. Separation of singing voice is currently under investigation.

Acknowledgments: This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-04-1-0027). We thank G. Hu and M. Wu for many useful discussions.

5. REFERENCES

[1] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, Vol. 15, pp. 1135–1150, 2004.

[2] J. Sundberg,"The acoustics of the singing voice," *Scientific American*, pp. 82–91, Mar. 1977.

[3] S.K. Shandilya and P. Rao, "Retrieving pitch of the singing voice in polyphonic audio," *National conference on communications*, NCC 2003, IIT Madras, 2003.

[4] R. Meddis and M. Hewitt, "A unitary model of pitch perception," *Journal of the Acoustical Society of America*, Vol. 102, pp. 1811–1820, 1997.

[5] M. Goto, "A predominant-F0 estimation method for polyphonic musical audio signals," *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, pp. 1085–1088, 2004.

[6] A.P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, pp. 204–816, 2003.

[7] M. Wu, D.L. 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.

[8] M.P. Cooke, *Modeling auditory processing and organization*, Cambridge Univ. Press, Cambridge, U.K., 1993.

[9] H. Helmholtz, *On the sensations of tone as a physiological basis for the theory of music*, A.J. Ellis, Ed., Dover, New York, 1863.

[10] P. Boersma, and D. Weenink, "Praat: Doing phonetics by computer, version 4.0.26" (http://www.fon.hum.uva.nl/praat), 2002.