MULTI-PITCH TRACKING USING GAUSSIAN MIXTURE MODEL WITH TIME VARYING PARAMETERS AND GRATING COMPRESSION TRANSFORM

Abhijith M N

Dept of Electrical Communication Engineering
Indian Institute of Science
Bangalore, India
mabhijithn@ece.iisc.ernet.in

Prasanta K Ghosh, K Rajgopal

Dept of Electrical Engineering
Indian Institute of Science
Bangalore, India
{prasantg,kasi}@ee.iisc.ernet.in

ABSTRACT

Grating Compression Transform (GCT) is a two-dimensional analysis of speech signal which has been shown to be effective in multi-pitch tracking in speech mixtures. Multi-pitch tracking methods using GCT apply Kalman filter framework to obtain pitch tracks which requires training of the filter parameters using true pitch tracks. We propose an unsupervised method for obtaining multiple pitch tracks. In the proposed method, multiple pitch tracks are modeled using time-varying means of a Gaussian mixture model (GMM), referred to as TVGMM. The TVGMM parameters are estimated using multiple pitch values at each frame in a given utterance obtained from different patches of the spectrogram using GCT. We evaluate the performance of the proposed method on all voiced speech mixtures as well as random speech mixtures having well separated and close pitch tracks. TVGMM achieves multi-pitch tracking with 51% and 53% multi-pitch estimates having error $\leq 20\%$ for random mixtures and all-voiced mixtures respectively. TVGMM also results in lower root mean squared error in pitch track estimation compared to that by Kalman filtering.

Index Terms— Grating Compression Transform, multi-pitch tracking, Gaussian mixture model, expectation- maximization

1. INTRODUCTION

Multi-pitch tracking in speech mixtures is a challenging problem which has seen considerable research over the years [1, 2, 3]. Popular multi-pitch tracking algorithms estimate pitches by analyzing each short-time frame of speech mixture signal. This is followed by statistical or probabilistic model to estimate individual pitch contours or pitch tracks. Wu, et. al, [4] used pitch period statistics and channel selection mechanism followed by hidden Markov model (HMM) based pitch contour generation. Jin and Wang [5] used correlogram and cross-channel correlation features followed by an HMM based pitch track generation. Recently, probabilistic models using factorial HMM (FHMM) [6] were used in obtaining multi-pitch tracks.

A class of approaches in multi-pitch tracking involves a two-dimensional analysis of speech signal proposed by Quatieri [7] and Ezzat, et. al, [8]. The 2-D analysis was supported by findings from the auditory neurophysiology [9] and is aimed at better understanding about the way phenomena including harmonicity, formants, onsets are observed in the spectrogram. The analysis was termed Grating Compression Transform (GCT) [7]. Quatieri and Wang [10] demonstrated the ability to extract multi-pitch estimates from GCT representation of small regions or patches of the spectrogram. In their work, pitch candidates were firstly obtained from the centroids of the pitch clusters from GCT of patches of the spectrograms. These pitch candidates were fed into a trained Kalman filter followed by assignment of estimated pitches to individual speaker's pitch contour and smoothing of pitch tracks. In contrast, we propose to model the pitch estimates, obtained from the GCT of patches of the spectrogram, as if they are generated from a Gaussian mixture model (GMM) with time-varying mean parameter, referred to as TVGMM. The means of the individual Gaussian components over time frames represent multiple pitch tracks. We develop a modified Expectation-Maximization (EM) algorithm to estimate time-varying parameters of the GMM. The proposed algorithm automatically handles assignment of pitch estimates in each frame to appropriate pitch tracks as well as ensures smooth pitch tracks. We evaluate the proposed multi-pitch tracking algorithm on all voiced and random speech mixtures with pitch tracks that are close as well as separated. TVGMM achieves multi-pitch tracking with 51% and 53% multi-pitch estimates having error $\leq 20\%$ for random mixtures and all-voiced mixtures respectively. In the case of average root mean squared error, TVGMM achieves an improvement by 13% and 6.4% over Kalman for all-voiced and random speech mixtures respectively.

The paper is organized as follows. Section 2 summarizes the GCT. The formulation of the proposed TVGMM is described in Section 3. Section 3.1 provides derivation of EM algorithm for estimating time-varying parameters of the TVGMM by jointly using GCT based multiple pitch estimates across all frames of an utterance. Section 4 describes the setup for multi-pitch tracking experiments and gives a quantitative evaluation of TVGMM.

2. GRATING COMPRESSION TRANSFORM

GCT is a framework for 2-D processing of speech in which 2-D analysis is performed on a time-frequency distribution of speech, viz. the narrowband spectrogram. GCT provides us an approach which helps in developing a modulation model for speech production, considering modulations along time as well as frequency [11]. The local 2-D Fourier transform of the narrowband spectrogram maps the harmonically-related signal components in the spectrogram into impulses in 2D Fourier plane - referred to as GCT plane [7]. Position of the impulses, its distance from the origin and orientation, helps us obtain the pitch (fundamental frequency) and pitch dynamics (rate of change of pitch) of the speaker respectively.

The harmonic line structure of the spectrogram over a small region or patch in a voiced segment can be modeled by a 2-D sinusoidal function sitting on a DC pedestal [11] .

$$s[n,m] = K + \alpha\cos(\Phi[n,m]) \tag{1}$$

where, $\Phi[n,m] = \omega_k\left(n\cos\Omega + m\sin\Omega\right) + \psi : \omega_k, \Omega, \psi, \alpha$ correspond to the frequency, orientation, phase, and amplitude of the 2-D sinusoid, respectively. The 2-D Fourier transform of the sequence in

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

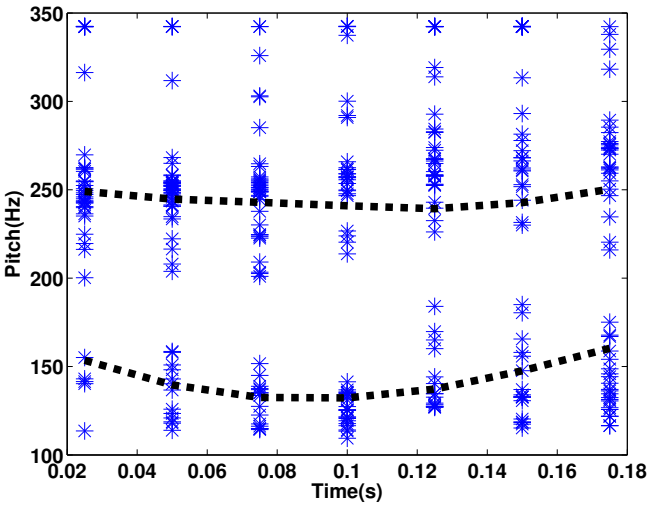

Fig. 1. Pitch estimates obtained from patches (stars) and true multi-pitch track(bold dashed lines) of a small extract from an all-voiced mixture of two speakers' speech.

eq. (1) is given by

$$S(\omega_1,\omega_2) = 2\pi\delta(\omega_1,\omega_2) + \alpha\pi\delta(\omega_1+\omega_k\sin\Omega,\omega_2-\omega_k\cos\Omega) + \alpha\pi\delta(\omega_1-\omega_k\sin\Omega,\omega_2+\omega_k\cos\Omega) \quad (2)$$

The 2-D Fourier transform given by eq. (2) consists of an impulse at the origin corresponding to the flat pedestal and impulses at $\pm(-\omega_k\sin\Omega,\omega_k\cos\Omega)$ corresponding to the sinusoid. The distance of the impulses from the origin along the ω_2 frequency axis represents the pitch, i.e., f_0, of the speaker for a. The pitch f_0 is given by [11]:

$$f_0 = \frac{2\pi f_s}{N_{STFT}\ \ \omega_k\cos\Omega} \quad (3)$$

where, f_s is sampling frequency of the speech signal, N_{STFT} is FFT length of the individual spectrogram or short-time Fourier transform (STFT) frames. In a speech signal with overlapped speech from two speakers, GCT could be applied on the overlapping voiced regions of both speakers. In this case, GCT of a patch of the spectrogram would show two distinct peaks or impulses. Thus estimates of multiple pitches- two in the mentioned case- can be obtained using eq. (3).

3. MULTI-PITCH TRACKING USING GCT AND TVGMM

As an illustration of the multi-pitch tracking, we consider an all-voiced mixture of two sentences from TIMIT corpus [12] "Where were you while we were away" and "He will allow a rare lie" spoken by a female and a male speaker respectively. The spectrogram of the mixture is obtained by performing 512 point FFT of an analysis window of 25ms duration with 90% overlap. The GCT is performed on the spectrogram patches of size 600Hz $\times$ 100ms at every 150Hz along the frequency axis and every 25ms along the time axis. Fig. 1 shows pitch estimates obtained for a small extract of this mixture.

Note that the pitch estimates obtained from the GCT of patches of the spectrogram occur in clusters corresponding to two speakers' pitch tracks. Quatieri and Wang [10] used the centroids of the clusters as pitch candidates of two speakers. In this paper we assume that the pitch estimates obtained from the spectrogram patches are generated from a GMM with time-varying means of individual Gaussian components corresponding to the individual pitch tracks at each frame. However, the challenge lies in associating pitch candidates from successive frames to obtain a pitch track. In [10], a pair of trained Kalman filters is used to obtain pitch tracks using observations as the cluster centroids in each frame. We assume that the means of the components of GMM to be time varying whose parameters are estimated jointly using pitch candidates over all time frames of an utterance. This automatically takes care of the association of pitch estimates across frames to each of the multiple pitch tracks. Moreover, we parameterize the time-varying means of TVGMM by a K-th degree polynomial, to ensure a smooth pitch track.

3.1. Estimation of TVGMM parameters

Consider the general case of a speech mixture consisting of J speakers where we have to estimate J pitch tracks. We assume that the candidate pitch estimates in the n-th frame are generated by a 1-D GMM with J components, with mean of each component corresponding to the true pitch of one of the J speakers. Hence, the distribution of pitch estimates in n-th frame is :

$$f_{\mathbf{x}_n}(x_{n,i}|\theta) = \sum_{j=1}^{J}\lambda_j(n)\phi\left(x_{n,i}|\theta_j(n)\right) \quad (4)$$

where, $\theta_j(n) = \{\mu_j(n),\sigma_j^2(n)\}$ are the mean and variance of the jth Gaussian component in frame n respectively, with $\mathbf{x}_n = \{x_{n,1},x_{n,2},...x_{n,L(n)}\}$ where $x_{n,i}$ is ith pitch estimate in n-th frame, $L(n)$ is the number of candidate pitch estimate samples in n-th frame and

$$\phi(x_{n,i}|\theta_j(n)) = \frac{1}{\sigma_j(n)\sqrt{2\pi}}\exp\left(-\frac{(x_{n,i}-\mu_j(n))^2}{2\sigma_j^2(n)}\right) \quad (5)$$

We model $\mu_j(n)$ by a K-th degree polynomial, to enforce smoothness of pitch tracks:

$$\mu_j(n) = \sum_{k=0}^{K} a_{k,j}n^k \quad (6)$$

For simplicity, we assume the variance and weight of individual Gaussian mixture components to be constant over time, i.e., $\sigma_j^2(n) = \sigma_j^2$ and $\lambda_j(n) = \lambda_j$. The parameters of TVGMM are estimated by the classical EM algorithm [13]. We estimate coefficients $a_{k,j}$, $k = 0,1,2,\cdots,K$ and $j = 1,2,\cdots,J$ in eq. (6) and thereby the time-varying means which are finally declared as the estimates of the pitch tracks, i.e., $\hat{f}_0(n,j) = \hat{\mu}_j(n)$where $\hat{f}_0(n,j)$ is the estimated pitch of the j-th track at the n-th frame.

In EM algorithm, a random variable $Z_{ij}(n)$ is introduced which indicates the component j which produces the sample $x_{n,i}$. The algorithm estimates parameters in two steps, viz. expectation (E - step) and maximization (M - step), performed iteratively till convergence

$$\underline{\mathrm{E-Step}}: \quad Q(\theta_j(n)|\theta_j^k(n)) = E_{\mathbf{Z}_n|\mathbf{x},\theta_j^k(n)}\ln\left(f\left(\mathbf{X},\mathbf{Z}_n|\theta_j(n)\right)\right) \quad (7)$$

$$\underline{\mathrm{M-step}}: \quad \theta_j^{(k+1)}(n) = \arg\max_{\theta_j(n)} Q(\theta_j(n)|\theta_j^k(n)) \quad (8)$$

where k is the iteration number, $\ln\left(f\left(\mathbf{X},\mathbf{Z}_n|\theta_j(n)\right)\right)$ is the data log likelihood with $\mathbf{Z}_n = Z_{ij}(n)$ $i = 1,2\cdots,L(n)$ $j = 1,2\cdots,J$ and $\mathbf{X} = [\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N]$ and N is number of frames under consideration. Since we are estimating parameters across multiple frames, the log likelihood is

$$\ln\left(f\left(\mathbf{X},\mathbf{Z}_n|\theta_j(n)\right)\right) = \sum_{n=1}^{N}\sum_{i=1}^{L(n)}\sum_{j=1}^{J} Z_{ij}(n)\ln\left(\lambda_j\phi\left(x_{n,i}|\theta_j(n)\right)\right) \quad (9)$$

Thus the E-Step of (7) reduces to,

$$Q(\theta_j(n)|\theta_j^k(n)) = \sum_{n=1}^{N}\sum_{i=1}^{L(n)}\sum_{j=1}^{J} \quad E_{\mathbf{Z}_n|\mathbf{x},\theta_j^k(n)}\left[Z_{ij}(n)|\mathbf{x},\theta_j^k(n)\right] \times \ln\left(\lambda_j\phi\left(x_{n,i}|\theta_j(n)\right)\right) \quad (10)$$

Now

$$\gamma_{ij}^k(n) = E_{\mathbf{Z}|\mathbf{x},\theta_j^k(n)}\left[Z_{ij}(n)|\mathbf{x},\theta_j^k(n)\right] = \frac{\lambda_j^k\phi\left(x_{n,i}|\theta_j^k(n)\right)}{\sum_{j=1}^{J}\lambda_j^k\phi\left(x_{n,i}|\theta_j^k(n)\right)} \quad (11)$$

Since we have assumed fixed variance σ_j^2 and component weights λ_j, the parameters to be maximized over in M-step of eq. (8) are the coefficients of the polynomial in eq. (6) denoted by,

$$\mathbf{a}_j = \{a_{k,j}\}_{k=0}^{K} \text{ with } j = 1, 2\cdots, J$$

Using eq. (5) eq. (10) and eq. (11), M-step of eq. (8) reduces to,

$$\mathbf{a}_j^{k+1} = \arg\max_{\mathbf{a}_j} Q(\mathbf{a}_j|\mathbf{a}_j^k) \quad (12)$$

$$\text{where, } Q(\mathbf{a}_j|\mathbf{a}_j^k) = \sum_{n=1}^{N}\sum_{i=1}^{L(n)}\sum_{j=1}^{J}\gamma_{ij}^k(n)\Big[\ln\lambda_j - \ln\left(\sigma_j\sqrt{2\pi}\right) - \frac{\left(x_{i,n} - \sum_{k=0}^{K}a_{k,j}n^k\right)^2}{2\sigma_j^2}\Big],$$

where σ_j^2 and λ_j are known constants. We proceed with the gradient descent algorithm to find the argument that maximizes eq. (12) and hence,

$$a_{l,j}^{(k+1)} = a_{l,j}^{(k)} + \eta\frac{\partial Q(\mathbf{a}_j|\mathbf{a}_j^k)}{\partial a_{l,j}} \quad l = 0, 1, 2\cdots, K \quad j = 1, 2\cdots, J \quad (13)$$

After simplification, the above equation yields

$$a_{l,j}^{(k+1)} = a_{l,j}^{(k)} + \eta\sum_{n=1}^{N}\sum_{i=1}^{L(n)}\frac{\gamma_{ij}^k(n)n^l}{2\sigma_j^2}\left(x_{i,n} - \sum_{k=0}^{K}a_{k,j}n^k\right)^2 \quad (14)$$

with $l = 0, 1, 2\cdots, K$ and η is the step-size in gradient descent.

4. EXPERIMENT AND EVALUATION

4.1. Dataset

We evaluate the performance of multi-pitch tracking using TVGMM on all-voiced speech mixtures and random speech mixtures. Random mixtures implies that individual speech signals in the mixture are not entirely voiced. The mixtures were generated using sentences from the GRID corpus [14] which contains 1000 different sentences spoken by each of 34 talkers (18 male, 16 female). A total of 150 random mixtures were generated. Among these 150 mixtures, 125 mixtures have 'separated' pitch tracks since they were mixtures of one male and one female subjects' speech signals. The remaining 25 mixtures have 'close' pitch tracks. These are generated by mixing utterances from two female speakers. The all-voiced mixtures were generated from a locally recorded data set containing 4 males and 4 females speaking the voiced utterances given in Table 1. 390 mixtures were generated from above data set. Among these 218 mixtures had 'separated' pitch tracks and 172 had 'close' pitch tracks. The ground truth pitch tracks of a speech mixture are obtained by computing pitch trajectories of the individual speech signals using Praat [15].

s1 - "We owe you a yo-yo"
s2 - "Nanny may know my meaning
s3 - "Where were you while we were away"
s4 - "a e i o u"

Table 1. All voiced utterances used in evaluation

Method of estimation	Total # frames	FPE count (5%≤ error ≤20%)	GPE count (≥20%)
Kalman	8010	30.5%	52.7%
TVGMM	8010	32.4%	48.8%

Table 2. Performance on 150 random mixtures (male and female voices in a mixture signal) generated from GRID corpus

4.2. Experiment Setup

To compute multi-pitch tracks from a speech mixture, it is required to find regions in the mixture where both speakers have voiced portions. It has been shown that the average duration of vowels, the predominant class of voiced utterances, is around 250ms for British English [16]. Hence we consider the value of N to be 10 which corresponds to 250ms duration of speech as frame period is 25ms. Thus all overlapping voiced segments with length at least 10 frames are used for our experiments. The candidate pitch values in each frame are computed using GCT of patches of the spectrogram as described in Section 3. These candidates are used in estimation of mean of the $J = 2$ components, because there are two speakers in each speech mixture. $K = 4$ degree polynomial is used to model time-varying mean of the proposed TVGMM. K-Means clustering is used to cluster the candidate pitch values in each frame into J=2 clusters and the cluster centroids are computed. Following assignment of centroids across frames, a 4 degree polynomial is fitted to each centroid trajectory. The assignment is carried out using minimum absolute difference between centroids of successive frames. The coefficients of this polynomial are used to initialize $\mathbf{a_j}$ in the gradient descent algorithm given by eq. (14). The gradient descent method is carried out for 200 iterations with step size $\eta = 0.1$. Variance σ_j^2 is empirically fixed at 50. Equal weights were assigned to two components in TVGMM because in each frame GCT gives equal number of pitch estimates for both speakers.

4.3. Results and Analysis

The performance of TVGMM is compared with that using Kalman filter framework [11] referred to as "Kalman". To train the Kalman filter, 20% of all-voiced speech mixtures were used [11]. Two types of measures are used to evaluate the performance of multi-pitch tracking. The first type of measure further consists of two measures which computes the percentage errors of the individual multi-pitch estimates at each frame with respect to the true-pitch of individual speakers. These are termed as Fine pitch error (FPE) where, $5\% \leq$ error $\leq 20\%$, and Gross pitch error (GPE), where, error $\geq 20\%$. The second type of measure is the root mean-squared error (RMSE) which is used to represent the error in estimating an entire pitch track. RMSE

Method of estimation	Total # frames	FPE count (5%≤ error ≤20%)	GPE count (≥20%)
Kalman	19340	32.8%	50.9%
TVGMM	19340	35.1%	46.7%

Table 3. Performance on all voiced mixtures - generated from the dataset using 4 male and 4 female speakers.

Type of Mixture	TVGMM	Kalman
All-voiced	75.8 (±32.7)	89.2 (±42.7)
Random	92.2 (±47.5)	98.5 (±47)

Table 4. Average RMSE in Hz obtained over all-voiced and random mixtures using TVGMM and Kalman methods of multi-pitch estimation. Standard deviation is written in brackets.

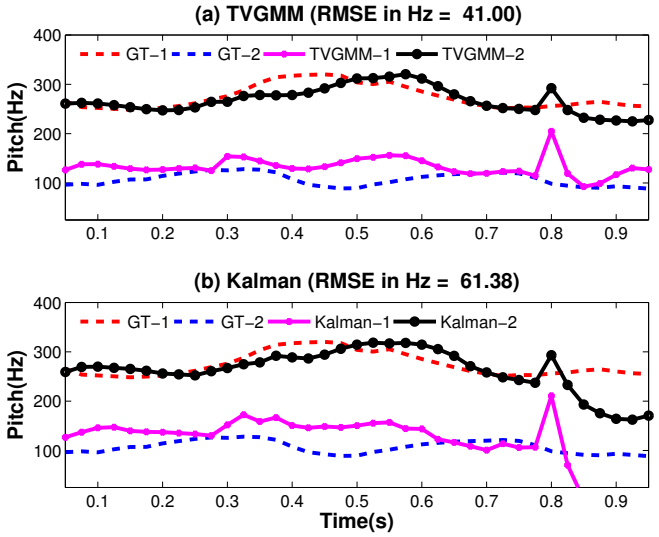

Fig. 2. Illustration of multi-pitch tracking on an all-voiced mixture –Ground truth(GT) pitch tracks and estimated pitch tracks, using (a) TVGMM method and (b) Kalman method.

(in Hz) is defined as [11]:

$$\text{RMSE} = \sqrt{\frac{1}{2N}\sum_{j=1}^{2}\sum_{n=1}^{N}\left(\hat{f}_0(n,j) - f_0(n,j)\right)^2} \qquad (15)$$

where N is number of frames over which pitch tracks are estimated. $\hat{f}_0(n,j)$ and $f_0(n,j)$ are estimated and true pitch respectively of jth speaker at n-th frame. Note that FPE and GPE are local frame based measures while RMSE is a global utterance level measures.

Tables 2 and 3 give comparison of the TVGMM and the Kalman using FPE and GPE for random mixtures and all-voiced mixtures respectively. Table 4 provides comparison of the methods based on average RMSE over all-voiced and random speech mixtures. Fig. 2 shows the estimated pitch tracks using TVGMM and Kalman methods along with true pitch tracks for an all-voiced speech mixture. Fig. 3 illustrates the same for a random speech mixture.

While using TVGMM, a 4% reduction in GPE over Kalman are observed for all-voiced speech mixtures and random mixtures. TVGMM method gives 51% of pitch estimates with error $\leq$ 20% for random mixtures. For all-voiced mixtures 53% of pitch estimates have error $\leq$ 20%. Kalman method is found to give a lower percentage of pitch estimates with error $\leq$ 20% for both all-voiced and random speech mixtures. Considering average RMSE, TVGMM achieves an improvement by 15% and 6.4% over Kalman for all-voiced and random speech mixtures respectively. We also applied state-of-the-art multi-pitch tracking algorithm proposed by Jin and Wang [5] on all-voiced speech mixtures and obtained 47% multi-pitch estimates with error $\leq$ 20% with an average RMSE of 138.2 Hz, which is higher compared to that by TVGMM. However, it should be noted that unlike TVGMM, the approach in [5] does not assume the locations of the voiced segments to be known.

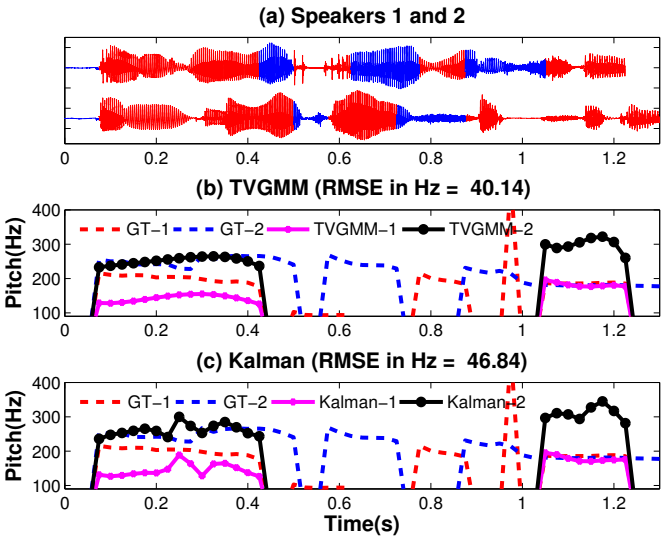

Fig. 3. (a) Speech waveforms of two speakers in a random mixture with red region showing where they are voiced.(b) and (c) show Ground truth(GT) pitch tracks and estimated pitch tracks,using TVGMM method and Kalman method respectively. Pitch tracks are plotted for overlapping voiced regions.

The above results show that TVGMM yields better multi-pitch estimates than Kalman. However, TVGMM uses a gradient descent algorithm to estimate its parameters whereas Kalman uses a trained filter to estimate multi-pitch tracks. Hence, TVGMM is slower in estimating multi-pitch tracks than Kalman[1]. But Kalman requires training which in turn requires speech mixtures with known pitch tracks. Hence, in the absence of a data set of speech mixtures with known multi-pitch values, TVGMM is suitable to obtain multi-pitch tracks. While Kalman could be useful for online estimation (because of faster processing), TVGMM can be used for offline estimation of multi-pitch tracks to achieve better estimation accuracy.

5. ACKNOWLEDGEMENT

We thank Prof. Chandra Sekhar S and Mr. Haricharan A of IISc, Bangalore for providing us with the all-voiced data set.

6. CONCLUSIONS

In this paper, we propose TVGMM for estimating multi-pitch tracks in speech mixtures. The proposed approach uses multiple candidate pitch values obtained using GCT followed by an estimation of time-varying means of TVGMM which provide estimates of individual pitch tracks. The method was applied on all-voiced as well as random two speaker speech mixtures. The results show that TVGMM performs better than Kalman filtering in the GCT based multi-pitch tracking. In the present formulation of TVGMM, the locations of the overlapping voiced segments are assumed to be known. Apriori detection of the overlapping voiced segments would make the TVGMM approach more robust. Multi-pitch track estimates could be improved by using additional pitch dynamics information from GCT. The proposed multi-pitch tracking method can also be extended to non-speech signals showing high degree of harmonicity like pitched instruments in music. These are parts of our future works.

[1]An Intel Core i7 processor running MATLAB R2013A, estimates TVGMM parameters and thereby multi-pitch tracks, of a speech mixture of duration 1.5s in 8s. Kalman method requires 0.01s for same. MATLAB code: `http://www.ee.iisc.ernet.in/new/people/faculty/prasantg/Softwares/tvgmm_alg.zip`

7. REFERENCES

[1] Anssi P Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 804–816, 2003.

[2] Mads Græsbøll Christensen and Andreas Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[3] Anssi Klapuri, "Number theoretical means of resolving a mixture of several harmonic sounds," , no. 200, pp. 400–403, 1998.

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

[5] Zhaozhang Jin and DeLiang Wang, "HMM -based multipitch tracking for noisy and reverberant speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1091–1102, 2011.

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

[7] Thomas F Quatieri, "2-d processing of speech with application to pitch estimation.," *Proc. INTERSPEECH*, pp. 1737–1740, 2002.

[8] Tony Ezzat, Jake V Bouvrie, and Tomaso Poggio, "Spectro-temporal analysis of speech using 2-d gabor filters.," in *Proc. INTERSPEECH*, 2007, pp. 506–509.

[9] Taishih Chi, Powen Ru, and Shihab A Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 2005.

[10] Tianyu T Wang and Thomas F Quatieri, "Multi-pitch estimation by a joint 2-d representation of pitch and pitch dynamics.," *Proc. INTERSPEECH*, pp. 645–648, 2010.

[11] Tianyu T Wang and Thomas F Quatieri, "Two-dimensional speech-signal modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1843–1856, 2012.

[12] "Timit acoustic-phonetic continuous speech corpus," `http://www.ldc.upenn.edu/Catalog/LDC93S1.html`.

[13] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[14] "The grid audiovisual sentence corpus," `http://spandh.dcs.shef.ac.uk/gridcorpus/`.

[15] Paul Boersma and David Weenink, "Praat:doing phonetics by computer," `http://www.fon.hum.uva.nl/praat/`.

[16] JG Wells, *A study of the formants of the pure vowels of British English*, University of London, 1962.