Robust Pitch Estimation and Tracking for Speakers Based on Subband Encoding and the Generalized Labeled Multi-Bernoulli Filter

Shoufeng Lin, *Member, IEEE*

***Abstract*—This paper proposes a new pitch estimator and a novel pitch tracker for speakers. We first decompose the sound signal into subbands using an auditory filterbank, assuming time-frequency sparsity of human speech. Instead of directly selecting the number of subbands according to experience, we propose a novel frequency coverage metric to derive the number of subbands and the center frequencies of the filterbank. The subband signals are then encoded inspired by the computational auditory scene analysis (CASA) approach, and the normalized autocorrelations are calculated for pitch estimation. To suppress spurious errors and track the speaker identity, the temporal continuity constraint is exploited and a Generalized Labeled Multi-Bernoulli (GLMB) filter is adapted for pitch tracking, where we use a novel pitch state transition model based on the Ornstein-Uhlenbeck process, and the measurement driven birth model for adaptive new births of pitch targets. Experimental evaluations with various additive noises demonstrate that the proposed methods have achieved better accuracy compared with several state-of-the-art pitch estimation methods in most studied scenarios. Tests using real recordings in a reverberant room also show that the proposed method is robust against reverberation.**

***Index Terms*—pitch tracking, auditory filterbank, CASA, frequency coverage, autocorrelation, GLMB tracking filter, Ornstein-Uhlenbeck process, measurement driven birth.**

I. Introduction

PITCH estimation and tracking can play an important part in many audio signal processing applications including automatic speaker identification, speech separation and transcription. In this paper, we focus on extracting fundamental frequencies[1] of human speakers, from single channel sound signals, which can be speech signals from a single speaker or concurrent speakers, mixed with noises or reverberation from the environment.

Many efforts have been made in estimating the fundamental frequency of voiced sound signals. Time domain methods investigate the periodic patterns of signals, and often apply the autocorrelation function (ACF), cross-correlation function (CCF), average magnitude difference function (AMDF) or the cumulative mean normalized difference function, etc. to the sound signals to detect the time delays that correspond to the fundamental periods [1]–[5]. Frequency domain methods study the harmonic structure of sound signal spectra and extract pitch information based on various features and rules, e.g. the harmonic product spectrum [6], subharmonic summation [7], wavelet based instantaneous frequency [8], [9] and the subharmonic-to-harmonic ratio [10], etc. Most of the existing methods can produce reliable pitch estimation results in amiable environments, but strong noises or reverberation can degrade the performance significantly, by corrupting the periodic patterns of time domain signals or the harmonic structures of the signal spectra. Other recent advances on robust pitch estimation include mainly the more complicated features and strategies, e.g. total energy of harmonics [11], harmonic frequency deviation [12], etc. The subspace-based method [13]–[15] have been developed to decouple speech and noise subspaces and can provide high resolution pitch estimates. Some more statistical methods provide probabilistic models for noisy sound signals and find pitch estimates with optimal probabilities according to their models and the observations [16], [17]. However, a majority of existing methods are designed for the pitch estimation of a single speaker.

For co-channel multi-pitch estimation of concurrent speakers, several works inspired by the computational auditory scene analysis (CASA) approaches (e.g. [18]) have been developed [19], [20]. They work on the time-frequency (TF) domain by decomposing single-channel sound signal into subbands via an auditory filterbank and then performing time-domain analysis in each subband. Although the center frequencies of subband filters of an auditory filterbank can be derived by selecting a number of subbands equidistantly on a frequency scale, e.g. the logarithmic scale [10], Bark scale [21] or the ERB-rate scale [20], [22], [23], etc., the selection of the total number of subbands has been essentially empirical. Apparently, using more subbands than necessary can impair computational efficiency, while insufficient subbands can cause loss of information and hence estimation errors. Furthermore, to obtain continuous pitch contours, temporal continuity constraint of pitch is often exploited, assuming continuous speech production by the human vocal system. Several pitch tracking methods based on the hidden Markov model (HMM) [12], [20], [24], [25] can be found in the literature, forming pitch tracks via estimating the hidden state sequence from observations. A recent work [26] uses trained Gaussian mixture models (GMMs) for signal spectrogram and put the probabilistic speaker observation models under the factorial hidden Markov models (FHMM) framework for multi-pitch tracking. Neural networks (NN) form another emerging approach for pitch estimation and tracking, using various features, neurons and network topologies [27]–[29]. However, for accurate pitch estimation and tracking results,

Shoufeng Lin is currently with the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, Western Australia. E-mail: shoufeng.lin@postgrad.curtin.edu.au; ee.linsf@gmail.com.

[1] In this paper, we use “pitch” and “fundamental frequency” interchangeably.

the HMM and NN based methods usually require carefully training the algorithms to obtain accurate hidden state transition probabilities, which can be inconvenient and restricting in practice.

In this paper, we first propose a novel speaker pitch estimator. It uses an auditory filterbank to decompose speech signals into subbands, based on the TF sparsity assumption [30] of speech signals. The number of subbands (and hence center frequencies) of the filterbank is calculated according to our proposed "frequency coverage" metric for consistent and full coverage over the frequency range. Moreover, inspired by the CASA approach and psychoacoustic studies, we propose to encode the subband signals with a robust encoding model to obtain distinct and reliable pitch estimates for the possibly noisy and quasi-periodic speech signals. Pitch estimates are then selected from the normalized autocorrelation coefficients of the encoded subband signals. Some preliminary results can be found in [31], and this paper provides detailed derivations and further evaluations.

We also propose a novel "training-free" pitch tracker based on the Generalized Labeled Multi-Bernoulli (GLMB) filter [32]–[34], to further reduce spurious estimation errors and to track pitch estimates with identities (i.e. to associate the pitch estimate with the corresponding speaker). The GLMB filter has been successful in tracking locations of multiple targets, but necessary adaptations are required for the pitch tracking problem. We propose to model the pitch state transition as an Ornstein Uhlenbeck process [35], assuming temporal continuity of speech production and that the pitch of a speaker tends to return to its average level. We also apply the measurement driven birth model for the adaptive new births of pitch targets in the GLMB prediction steps [33], [36], and provide the adaptations for the cases of a single pitch target and long pauses during speech. The resulting pitch tracking filter also assigns a unique identity to pitch estimates of a corresponding speaker, and can thus form linked tracks of pitch estimates for the respective speakers over time. This novel pitch tracker is applicable to tracking pitches of a single speaker, as well as concurrent speakers with pitches at different levels. It uses basic generic statistic models for the pitch state transition and observations, and therefore does not require training.

The rest of the paper is organized as follows. Section II presents our novel pitch estimator, which is followed by the proposed pitch tracker in Section III. Numerical studies are carried out in Section IV, and conclusions are provided in Section V.

II. Speaker Pitch Estimator

A. Speech Signal Model

In a noisy and reverberant environment, sound signal acquired by a single microphone is a mixture of reverberated speech signal from the speaker(s) and noise:

$$x(t) = \sum_{q=1}^{Q} s_q(t) * \mathrm{h}_q(t) + n(t), \tag{1}$$

where $t \in \mathbb{R}$ is the continuous time, the convolution operation is denoted as $*$, the additive noise at the microphone as $n(t)$. $\mathrm{h}_q(t) \in \mathbb{R}$ is the acoustic room impulse response (RIR) from the q-th speaker to the microphone. $q = 1, ..., Q$, and integer $Q \geq 1$ the number of concurrent speakers. $s_q(t)$ is the sound signal from speaker q. The unvoiced part of $s_q(t)$ is often regarded as a stochastic process, while the voiced part can be modeled based on the source excitation - vocal tract models for the process of speech production [37], and the amplitude-modulation (AM) and frequency modulation (FM) structure [38], [39]:

$$s_q(t) = \sum_{\hbar=1}^{H_q} A_q^{(\hbar)}(t) \cdot \cos\big(2\pi \cdot \hbar \cdot f_q \cdot t + \phi_q^{(\hbar)}(t)\big), \tag{2}$$

where integer $\hbar$ the order of harmonics for a speaker, integer H_q the maximum order of harmonics for speaker q, $A_q^{(\hbar)}(t) \geq 0$ the envelope of each harmonic, $\phi_q^{(\hbar)}(t) \in \mathbb{R}$ the slow time-varying phase (which makes the speech signals quasi-periodic), $f_q > 0$ the desired fundamental frequency. Compared to the modulating harmonic frequency, the envelope $A_q^{(\hbar)}(t)$ is usually narrow band. Note that the amplitude of the fundamental frequency component may not be the strongest, due to the speech production process.

B. Subband Decomposition

Based on the TF sparsity assumption [30], to separate harmonic components from the speaker(s), the microphone signal can be decomposed via an auditory filterbank [18], [39]:

$$x^{(b)}(t) = x(t) * g^{(b)}(t), \tag{3}$$

where $x^{(b)}(t)$ denotes the decomposed signals from the microphone in subband b, $b = 1, \ldots, N_b$, integer N_b is the total number of subbands, and $g^{(b)}(t)$ is the filter impulse response of subband b, which is aligned in time between subbands. Common auditory filters include the gammatone filter [18], [40], [41], gammachirp filter, etc. as well as their variants. In this paper, we use the gammatone filter in [41], which can be expressed as

$$g^{(b)}(t) = \tilde{g}^{(b)}(t) \cdot \cos(2\pi f_C^{(b)} t), \tag{4}$$

where

$$\tilde{g}^{(b)}(t) = (t + t_d)^{\vartheta - 1} e^{-2\pi f_b^{(b)}(t + t_d)}, \tag{5}$$

integer ϑ is the order of filter ($\vartheta = 4$ in this paper), t_d is time delay for alignment between filter bands, $f_b^{(b)}$ scaling factor for the bandwidth [18], [40], and $f_C^{(b)}$ is the center frequency of filter band b.

From (1), (2), (3), when the harmonic component $\hbar$ of the q-th speaker falls within the passband of subband b, and the noise is small in the particular subband, using the commutativity and associativity properties of convolution, and the frequency selectivity of the filterbank, the decomposed signals in subband b become:

$$\begin{aligned} x^{(b)}(t) &= \big[\sum_{q=1}^{Q} s_q(t) * \mathrm{h}_q(t) + n(t)\big] * g^{(b)}(t) \\ &\approx \sum_{q=1}^{Q} s_q(t) * \mathrm{h}_q(t) * g^{(b)}(t) \end{aligned} \tag{6}$$

Assuming that the reverberation is not too strong, the RIR can be simplified as

$$\mathrm{h}_q(t) = \mathrm{h}_q(t_{d_q}) \cdot \delta(t - t_{d_q}), \tag{7}$$

where $\mathrm{h}_q(t_{d_q}) > 0$ and $t_{d_q} > 0$ are the direct path amplitude and time-delay of the RIR from speaker q to the microphone.

Hence from (4), (6) and (7), when the harmonic frequency falls in the subband, we have (see Appendix A)

$$x^{(b)}(t) \approx \tilde{S}_q^{(b)}(t) \cdot \cos(\tilde{\phi}_q^{(b)}(t)), \tag{8}$$

where

$$\tilde{S}_q^{(b)}(t) = \frac{1}{2} \cdot \mathrm{h}_q(t_{d_q}) \cdot A_q^{(\hbar)}(t - t_{d_q}) * \tilde{g}^{(b)}(t), \tag{9}$$

$$\tilde{\phi}_q^{(b)}(t) = 2\pi\hbar f_q \cdot (t - t_{d_q}) + \phi_q^{(\hbar)}(t - t_{d_q}). \tag{10}$$

C. Frequency Range, Scale and Coverage

A critical part of the subband approach in speech processing is the selection of center frequencies $f_C^{(b)}$ for subband filters $g^{(b)}(t)$, according to the chosen frequency range $[f_{min}, f_{max}]$, where $f_{max} > f_{min} > 0$. This is usually addressed by choosing a frequency scale and the corresponding number of subbands N_b. Various frequency scales have been used in the pitch estimation literature, including the logarithmic [10], Bark [21] and ERB-rate scales [22]. However, in the current literature, the number of subbands for a given frequency scale in the given frequency range largely varies from one implementation to another, with no clear reason other than as an empirical choice. In [23], a total of 20 subbands are used for frequency range of 330Hz to 3700Hz, while [20] implements 128 gammatone filters between 80Hz and 5000Hz.

In this paper, we use the ERB-rate scale (ERBS) as developed in [22]. Denote the general form of ERB as

$$\upsilon(f) = D + E \cdot f, \tag{11}$$

where $D = 24.7$, and $E = 0.108$ as given in [22].

From (11), the resulting ERBS becomes (see Appendix B):

$$\Upsilon(f) \triangleq \int \frac{1}{\upsilon(f)} df = E' \lg(1 + D' \cdot f), \tag{12}$$

with the boundary condition $\Upsilon(0) = 0$. Here $D' \triangleq \frac{E}{D}$ and $E' \triangleq \frac{1}{E \cdot \lg e}$. As given in [22], $E' = 21.4$, and $D' = 0.00437$.

To derive the total number of subbands and the subband center frequencies for a given frequency range, we propose to use the "frequency coverage" metric, i.e.

$$\eta_C^{(b)} \triangleq \frac{\frac{1}{2} \cdot (f_B^{(b+1)} + f_B^{(b)})}{f_C^{(b+1)} - f_C^{(b)}}, \tag{13}$$

where $f_B^{(b)}$ denotes the filter bandwidth of subband b. As its name indicates, the "frequency coverage" metric measures how much of the frequency range is 'covered' by all the passbands of subband filters. This is easy to understand by first considering an ideal "brick-wall" bandpass filterbank. Fig. 1 also shows the intuition for the proposed frequency coverage metric using the Gammatone filter. For $\eta_C = 1$, the -3dB passbands of adjacent Gammatone filters align with no overlap. Apparently, a filterbank has consistent and full frequency coverage when $\eta_C^{(b)} \equiv 1$. For $\eta_C^{(b)} < 1$, there are some frequencies falling out of the passbands of the filterbank, which may result in estimation error when these frequencies include the desired fundamental frequency. The case of $\eta_C^{(b)} > 1$ still leads to full frequency coverage, but there are redundancies as some frequency components are captured and analysed multiple times.

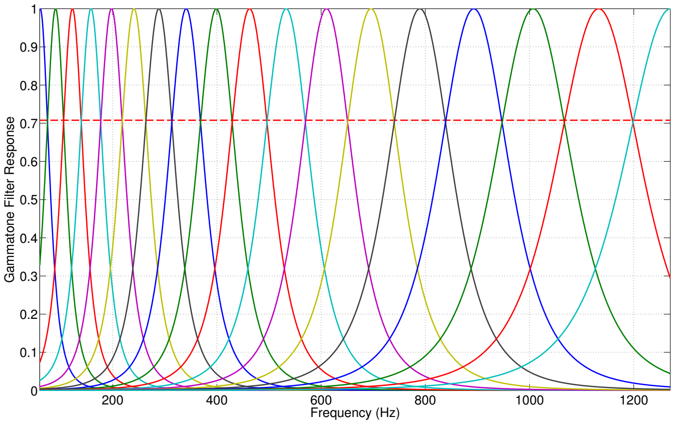

Fig. 1. An example of the frequency coverage metric (using the Gammatone filters). The frequency range is 60Hz to 1270Hz, thus for $\eta_c = 1$ there are 18 subbands and the -3dB passbands of resulting subband filters align.

The linear relationship between bandwidth and center frequency holds for certain types of filters. Particularly, for the gammatone filter we have [41]:

$$f_B^{(b)} = K_\vartheta \cdot f_b^{(b)} = K_\vartheta \cdot \upsilon(f_C^{(b)}), \tag{14}$$

where K_ϑ is a constant for a given filter order ϑ as given in (64). In particular, $K_4 = 0.887$. Here $\vartheta = 4$ is chosen for sufficient subband frequency selectivity (the attenuation is larger than 24dB at $f_C^{(b)} \pm 2f_B^{(b)}$ for the 4th-order gammatone filter of subband b).

The subband center frequencies in the given frequency range are distributed equidistantly on the ERBS, i.e.:

$$f_C^{(b)} = \Upsilon^{-1}\Big(\frac{(N_b - b) \cdot \Upsilon(f_{min}) + (b-1) \cdot \Upsilon(f_{max})}{N_b - 1}\Big). \tag{15}$$

Therefore the number of subbands N_b can be derived from (13), (14) and (15) (see Appendix C):

$$N_b = \mathrm{round}\Bigg(1 + \frac{\ln(\frac{D + E \cdot f_{max}}{D + E \cdot f_{min}})}{\ln(\frac{2\eta_C^{(b)} + E \cdot K_\vartheta}{2\eta_C^{(b)} - E \cdot K_\vartheta})}\Bigg). \tag{16}$$

This provides a consistent way for calculating the number of subbands in a given frequency range based on the frequency coverage metric. Once N_b is obtained, the center frequencies can also be calculated from (15). In this paper, we choose $\eta_C^{(b)} \equiv 1$ for full frequency coverage without redundancies in processing. Since we keep $\eta_C^{(b)}$ the same for all subbands, η_C is used hereafter for simplicity of denotation.

The pitch frequency range is denoted as $[F0_{min}, F0_{max}]$. In this paper, we choose $F0_{min} = 60$Hz, and $F0_{max} = 500$Hz to cover the pitch range of most speakers [37], [42]. Accordingly, the minimum subband frequency is chosen as

$f_{min} = F0_{min} = 60$Hz. It has been pointed out that while low frequency auditory nerve fibers of inner hair cells tend to phase lock to pitch stimulus, those of frequencies above 1300Hz do not [23]. Thus we choose $f_{max} = 1270$Hz in this paper [23]. Although autocorrelations of subband envelopes in frequencies higher than 1300Hz were used in [23], this high frequency range is not needed in our proposed method. Thus for $\eta_C = 1$ we can get $N_b = 18$ from (16) for the frequency range of $[60, 1270]$Hz.

D. Rectification and Pitch Encoding

In practice, signals are discretized at a sampling frequency of $f_s > 0$. We first half-wave rectify the discrete subband signal as in [3], [19], [43]:

$$\hat{x}^{(b)}(k/f_s) = \frac{1}{2}\cdot(x^{(b)}(k/f_s) + |x^{(b)}(k/f_s)|), \tag{17}$$

where the discrete time index $k \in \mathbb{Z}$.

Assuming a slow-changing $\phi_q^{(\hbar)}(t)$ in (8), we can rewrite the half-wave rectified subband signal as a convolution:

$$\hat{x}^{(b)}(k/f_s) \approx \zeta_{\text{cosine}}^{(\hbar,q)}(k) * \sum_{\hat{k}_n^{(b)} \in \hat{K}^{(b)}} \tilde{S}_q^{(b)}(k/f_s)\cdot\delta(k - \hat{k}_n^{(b)}), \tag{18}$$

where $\delta(\cdot)$ is the Dirac delta function, and $\zeta_{\text{cosine}}^{(\hbar,q)}(k)$ is the non-negative part of the cosine term with peak at $k = 0$, i.e.

$$\zeta_{\text{cosine}}^{(\hbar,q)}(k) \triangleq \cos(2\pi\hbar f_q \cdot k/f_s), k \in [-\frac{f_s}{4\hbar f_q}, \frac{f_s}{4\hbar f_q}], \tag{19}$$

$\hat{K}^{(b)} \triangleq \{\hat{k}_n^{(b)} |\ n = 0, 1, ...\}$, and $\hat{k}_n^{(b)}$ is the index of a local peak

$$\hat{k}_n^{(b)} = \arg\max_k \hat{x}^{(b)}(k/f_s),\ \forall\ k \in (k_{n-}^{(b)}, k_{n+}^{(b)}), \tag{20}$$

$k_{n-}^{(b)}, k_{n+}^{(b)}$ are consecutive zero-crossings of $\hat{x}^{(b)}(k/f_s)$ that satisfy

$$\hat{x}^{(b)}(k/f_s) > 0,\ \forall\ k \in (k_{n-}^{(b)}, k_{n+}^{(b)}). \tag{21}$$

The speaker pitch can be found from the periodicity information of scaled delta functions, by searching for the peak of autocorrelation results, but the slow-changing cosine term can make the peak widespread or even cause spurious estimates. Actually we can check the time intervals between peaks of the scaled delta functions, i.e. $\tilde{S}_q^{(b)}(k/f_s)\cdot\delta(k - \hat{k}_n^{(b)})$ as in (18). The problem is that the voiced speech signal is quasi-periodic, and the scaled delta functions alone can be sensitive to noise (the noise can affect the time indices of peaks), in the autocorrelation. Therefore, inspired by the approaches of computational auditory scene analysis (CASA) [18], [44], we propose to encode the subband signals as convolution of the scaled delta functions with a symmetrical encoding template, which in effect replaces the cosine term in (18):

$$\zeta_p^{(\hbar,q)}(k) \triangleq \begin{cases} e^{-|k|},\ k \in [-5, 5] \\ 0,\ \text{otherwise}, \end{cases} \tag{22}$$

where we empirically choose a fixed decay rate, so that the spike decays to 5% of its peak in about 0.2ms at a sampling rate of $f_s = 16000$Hz in this paper. This aligns with the psychoacoustic observation of the exponential decay of the synaptic cleft contents from the hair cell in the organ of Corti [44]. The encoding template is symmetrical to avoid bias of time delay estimation in the autocorrelation. Moreover, the encoding template can also be connected with the observation of the Laplacian distribution of peaks versus the relative time lags [17], [20], except that for simplicity we discard (truncate) smaller values in (22) and assume that its dependence on subband indices and speakers is negligible and the constant coefficient for the exponential term is 1 as it does not affect the resulting normalized correlation coefficients.

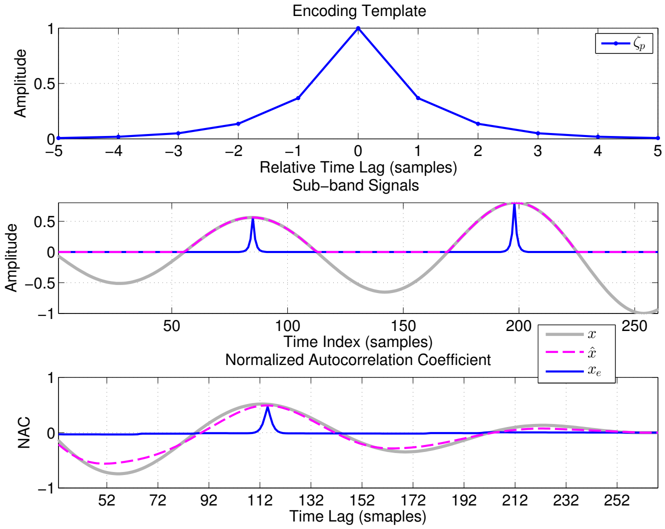

Fig. 2. Pitch encoding template (top panel), a subband signal from the filterbank, $x^{(b)}$, its half-wave rectified $\hat{x}^{(b)}$ and encoded signal (middle panel) $x_e^{(b)}$, and normalized autocorrelation coefficient of respective signals (bottom panel).

The resulting encoded subband signal from (18) and (22) is

$$x_e^{(b)}(k) = \zeta_p^{(\hbar,q)}(k) * \sum_{\hat{k}_n^{(b)} \in \hat{K}^{(b)}} \tilde{S}_q^{(b)}(k/f_s)\cdot\delta(k - \hat{k}_n^{(b)}). \tag{23}$$

The top two panels of Fig. 2 depict the encoding template, a segment of a subband signal $x^{(b)}$, its half-wave rectified signal $\hat{x}^{(b)}$ and its encoded signal $x_e^{(b)}$ respectively. Normalized autocorrelation coefficients of respective signal are plotted in the bottom panel, and to be discussed next.

E. Subband Autocorrelation and Pitch Extraction

The encoded subband signals are further processed via autocorrelation in frames of length $n_{corr} = \lceil 2\cdot f_s/F0_{min}\rceil$ and in step size of $n_{step} \in \mathbb{N}$. The range of sample delays is $d_\tau \in [d_{min}, d_{max}]$, where $d_{min} = \lfloor f_s/F0_{max}\rfloor$, $d_{max} = \lceil f_s/F0_{min}\rceil$. Here $\lfloor\cdot\rfloor$ denotes the largest integer less than or equal to a given number, while $\lceil\cdot\rceil$ denotes the smallest integer greater than or equal to a given number.

Normalized autocorrelation coefficients (NAC) for encoded subband b in the jth frame can be calculated using

$$A^{(b)}(j, d_\tau) = \frac{\sum_{k=(j-1)\cdot n_{step}+1}^{(j-1)\cdot n_{step}+n_{corr}-d_\tau} \tilde{x}_e^{(b)}(k)\cdot\tilde{x}_e^{(b)}(k+d_\tau)}{\sum_{k=(j-1)\cdot n_{step}+1}^{(j-1)\cdot n_{step}+n_{corr}} [\tilde{x}_e^{(b)}(k)]^2}, \tag{24}$$

where

$$\tilde{x}_e^{(b)}(k) = x_e^{(b)}(k) - \frac{1}{n_{corr}} \cdot \sum_{k'=(j-1)\cdot n_{step}+1}^{(j-1)\cdot n_{step}+n_{corr}} x_e^{(b)}(k'). \quad (25)$$

Compared with the cross-correlation function (see e.g. [2]) for pitch estimation, this autocorrelation method results in a decreasing envelope as the time delay increases, due to the decreasing length of data in the numerator, which actually helps in suppressing the sub-harmonic errors. Similarly, the subband signal and the half-wave rectified subband signal can be used instead of the encoded subband signal in (24) to calculate their corresponding NACs, and the results are given in the bottom panel of Fig. 2. We can see in this case that compared to the other two curves, the proposed subband encoding method produces a sharp peak corresponding to the expected period in the NAC, and there is no significant second peak in the expected range of sample delays $[d_{min}, d_{max}]$.

In each time frame, we use the average of the NAC over subbands:

$$A_{\sum}(j, d_\tau) = \frac{1}{N_b} \sum_{b=1}^{N_b} A^{(b)}(j, d_\tau). \quad (26)$$

Then the pitch(es) in each frame can be estimated from the sample delays that correspond to the peaks in $A_{\sum}(j, \cdot)$. The strongest peak over the threshold $T_{A_{\sum}} = 0.125$ (i.e. -9dB) is directly used for the single pitch estimation (cf. the correlogram in Fig. 3 for the selection of this threshold, which is found consistent over a range of test cases). Due to the quasi-periodic nature of speech signals, for multi-pitch estimation, weaker peaks at sample delays that correspond to harmonics or sub-harmonics of the stronger peaks are removed. The pseudocode of the pitch extraction for frame j is summarized in Algorithm 1.

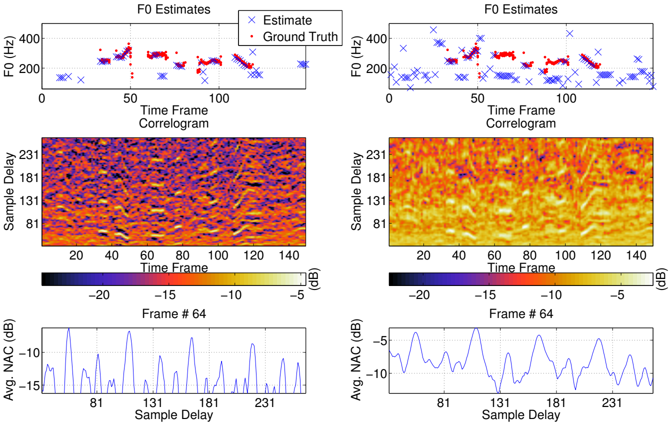

Fig. 3. Pitch estimation results (female speech with babble noise, SNR=5dB). Left column gives the pitch estimation results from proposed method. Right column shows the pitch estimation results using the autocorrelation of raw subband signals.

Algorithm 1 Pitch Extraction for the Pitch Estimator

```
Input: normalized autocorrelation coefficients $A_{\sum}(j, \cdot)$;
Output: pitch estimates $\widehat{F_0}\{j\}$.
procedure PITCH EXTRACTION
    Find Peaks:
    Find and Sort all peaks over a threshold, i.e. $A_{\sum}(j, \hat{d}_{\tau_i}) \geq T_{A_{\sum}}$, $i = 1, ..., N_k$, and $A_{\sum}(j, \hat{d}_{\tau_1}) \geq A_{\sum}(j, \hat{d}_{\tau_2}) \geq \cdots$;
    Pitch Extraction:
    if $N_k == 0$ then
        $\widehat{F_0}\{j\} = \emptyset$ (e.g. unvoiced sound segment or silence or miss-detection);
    else if $N_k \geq 1$ then
        if single pitch then
            $$\widehat{F_0}\{j\} = \{f_s/\hat{d}_{\tau_1}\}. \quad (27)$$
        else if multi-pitch then
            for $i = 1 : N_k$ do
                for $\rho = i : N_k$ do
                    if $\hat{d}_{\tau_\rho}$ is a harmonic or sub-harmonic of $\hat{d}_{\tau_i}$, then Discard $\hat{d}_{\tau_\rho}$.
                    end if
                end for
            end for
            $\widehat{F_0}\{j\} = \{\hat{f}_i \mid \hat{f}_i = f_s/\hat{d}_{\tau_i} \ \wedge \ \hat{f}_i \neq f_s/\hat{d}_{\tau_\rho}\}$.
        end if
    end if
end procedure
```

Fig. 3 provides a single pitch example comparing the proposed estimator (26) using the encoded subband signals (23) and using raw subband signals (3). The top row provides the resulted pitch estimation results using (27) from the proposed method and the reference method. We can see that the proposed method produces more valid estimates, while the reference method produces considerably more errors. The middle row depicts the correlogram (26) from the proposed method and the reference method. The proposed method produces more distinct pitch patterns. The bottom row shows the averaged autocorrelation results at frame 64, where the proposed method correctly produces the pitch estimate, while the reference method produces a sub-harmonic error. Therefore it is clear that the proposed method has distinct peaks by virtue of the proposed pitch encoding, while the peaks of the reference method are comparatively widespread. Moreover, using the raw subband signals produces more harmonics or sub-harmonics errors. For both cases, spurious estimates when there are no voiced sounds in the ground truth speech signal are from the babble noise.

As also can be seen from the top panel of Fig. 3, the pitch estimates $\widehat{F_0}\{j\}$ from (27) contains in most cases the desired pitch estimates compared with the ground truth, but occasionally there may still be the sub-harmonics, harmonics or other spurious errors, which do not form continuous pitch contours with neighbouring estimates. It can also be an empty set, especially in the case of unvoiced sounds or silence.

III. Pitch Tracker

In order to extract the desired pitch estimates of the speaker from $\widehat{F_0}\{j\}$, while suppressing the spurious errors (e.g. the pitch estimates that jump far away from pitch contours as shown in Fig. 3), we exploit the temporal continuity constraint for pitch contour assuming continuous speech production by the human vocal system. Further assuming that the pitch of a speaker tend to return to its average level, we propose to model the pitch transition with the Ornstein Uhlenbeck process [35]. For concurrent speakers, pitch tracking also aim at forming separate tracks of pitch estimates for respective speakers. However, more prior information (e.g. by training the algorithms) is usually required to separate pitches that overlap (i.e. when speakers have close levels of pitches). Nonetheless, we point out that in the case where pitches of concurrent speakers are at different levels, it is possible to track concurrent speaker pitches without the effort of training the algorithms. Thus we propose to treat the speaker as a target that has labeled (i.e. with identity) states (i.e. pitches) evolving over time, thereby tracking the pitch of speakers based on pitch observations (i.e. estimates $\widehat{F_0}\{j\}$ from Section II) and the GLMB [32]–[34] [2] online tracking framework. We also implement a measurement driven birth model [33], [36] for the practical adaptive pitch target births in the GLMB recursion, and present adaptations of the GLMB filter for the pitch tracking problem. In contrast to existing multi-pitch tracking methods [20], [26], the proposed method does not require training, as the models used do not rely on the speech database.

A. GLMB RFS Definitions

Denote the labeled state of pitch target as $\mathbf{x}_i \triangleq (\mathrm{x}_i, \ell_i) \in \mathbf{X}$, where i is index, x_i denotes the pitch state, and ℓ_i its label (target identity). The GLMB RFS $\mathbf{X} \triangleq \{(\mathrm{x}_i, \ell_i) \mid i \in \mathbb{N}\}$ is a labeled RFS [3] with state space $\mathbb{X}, (\mathrm{x}_i \in \mathbb{X})$ and label space $\mathbb{L}, (\ell_i \in \mathbb{L})$, where the labels are unique, i.e. $\ell_i \neq \ell_{i'}, \ \forall i \neq i'$. Its probability density is given as [32]

$$\pi(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{\xi \in \Xi} w^{(\xi)}(\mathcal{L}(\mathbf{X})) \left[p^{(\xi)}\right]^{\mathbf{X}}, \tag{28}$$

where the discrete index space Ξ is the space of association map histories. Each $\xi \in \Xi$ represents a history of association map up to current time. Each $p^{(\xi)}(\cdot, \ell)$ is the probability density of the states of target $\ell \in I = \mathcal{L}(\mathbf{X})$, and each $w^{(\xi)}(I)$ is non-negative with $\sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} w^{(\xi)}(I) = 1$. Projection $\mathcal{L} : \mathbb{X}\times\mathbb{L} \rightarrow \mathbb{L}$ is defined as $\mathcal{L}((\mathrm{x}, \ell)) = \ell$, and $\mathcal{L}(\mathbf{X}) = \{\mathcal{L}(\mathbf{x}) \mid \mathbf{x} \in \mathbf{X}\}$. $\mathcal{F}(\mathbb{X})$ denotes the class of finite subsets of a space $\mathbb{X}$. The function $\Delta(\mathbf{X}) \triangleq \delta_{|\mathbf{X}|}(|\mathcal{L}(\mathbf{X})|)$ is called the *distinct label indicator*, where $|\cdot|$ denotes the cardinality of an RFS. The RFS exponential notation is defined as $h^X \triangleq \prod_{\mathrm{x}\in X} h(\mathrm{x})$, where h is a real-valued function, with $h^{\emptyset} = 1$ by convention.

The δ-GLMB form of (28) is completely characterized by the set of parameters $\{(\omega^{(I,\xi)}, p^{(\xi)}) \mid (I, \xi) \in \mathcal{F}(\mathbb{L})\times\Xi\}$, with the probability density given as [34]

$$\pi(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \omega^{(I,\xi)} \delta_I(\mathcal{L}(\mathbf{X})) \left[p^{(\xi)}\right]^{\mathbf{X}}, \tag{29}$$

where the pair $(I, \xi) \in \mathcal{F}(\mathbb{L}) \times \Xi$ is called a *hypothesis*, and its associated weight $\omega^{(I,\xi)}$ the probability of the hypothesis.

$$\delta_Y(X) \triangleq \begin{cases} 1, \text{ if } X = Y \\ 0, \text{ otherwise.} \end{cases}$$

The GLMB recursion consists of the "update" step based on Bayes inference and the Chapman-Kolmogorov [35] "prediction" step based on the state transition model.

B. GLMB Recursion: Update

If the current RFS prediction density is a δ-GLMB of the form (29), using the current observation (pitch estimates) by denoting $\hat{F}_0 \triangleq \widehat{F_0}\{j\}$ as given in Algorithm 1, the posterior density is a δ-GLMB [34], i.e.

$$\pi(\mathbf{X}|\hat{F}_0) = \\ \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \ \sum_{\theta\in\Theta(I)} \omega^{(I,\xi,\theta)}(\hat{F}_0)\delta_I(\mathcal{L}(\mathbf{X}))\left[p^{(\xi,\theta)}(\cdot|\hat{F}_0)\right]^{\mathbf{X}}, \tag{30}$$

where $\Theta(I)$ denotes the subset of current association maps with domain I,

$$\omega^{(I,\xi,\theta)}(\hat{F}_0) \ \propto \ \omega^{(I,\xi)}[\eta_{F_0}^{(\xi,\theta)}]^I \tag{31}$$

$$p^{(\xi,\theta)}(\mathrm{x}, \ell|\hat{F}_0) \ = \ \frac{p^{(\xi)}(\mathrm{x}, \ell)\psi_{F_0}(\mathrm{x}, \ell; \theta)}{\eta_{F_0}^{(\xi,\theta)}(\ell)} \tag{32}$$

$$\psi_{F_0}(\mathrm{x}, \ell; \theta) \ = \begin{cases} \frac{p_D(\mathrm{x},\ell)\mathrm{g}(\hat{f}_{\theta(\ell)}|\mathrm{x},\ell)}{\kappa(\hat{f}_{\theta(\ell)})}, \text{if } \theta(\ell) > 0 \\ 1 - p_D(\mathrm{x}, \ell), \text{if } \theta(\ell) = 0 \end{cases} \tag{33}$$

$$\eta_{F_0}^{(\xi,\theta)}(\ell) \ = \ \left\langle p^{(\xi)}(\cdot, \ell), \psi_{F_0}(\cdot, \ell; \theta)\right\rangle \tag{34}$$

$\mathrm{g}(\hat{f}_{\theta(\ell)}|\mathrm{x}, \ell)$ is the likelihood for the measurement $\hat{f}_{\theta(\ell)} \in \hat{F}_0$ being generated by (x, ℓ), and $\kappa(\cdot)$ is the intensity function of Poisson RFS which we use to describe the clutter. p_D is the probability of a target state being detected. The standard inner product notation is defined as $\langle f, g\rangle \triangleq \int f(\mathrm{x})g(\mathrm{x})d\mathrm{x}$.

C. GLMB Recursion: Prediction

If the current RFS filtering density from its previous update step is a δ-GLMB of the form (29), the prediction density to the next time is a δ-GLMB given as [34]

$$\pi_+(\mathbf{X}_+) = \\ \Delta(\mathbf{X}_+) \sum_{(I_+,\xi)\in\mathcal{F}(\mathbb{L}_+)\times\Xi} \omega_+^{(I_+,\xi)} \delta_{I_+}(\mathcal{L}(\mathbf{X}_+)) \left[p_+^{(\xi)}\right]^{\mathbf{X}_+}, \tag{35}$$

where

$$\omega_+^{(I_+,\xi)} \ = \ \omega_S^{(\xi)}(I_+ \cap \mathbb{L}) w_B(I_+ \cap \mathbb{B}) \tag{36}$$

$$p_+^{(\xi)}(\mathrm{x}, \ell) \ = \ 1_{\mathbb{L}}(\ell)p_S^{(\xi)}(\mathrm{x}, \ell) + 1_{\mathbb{B}}(\ell)p_B(\mathrm{x}, \ell) \tag{37}$$

$$\omega_S^{(\xi)}(L) \ = \ [\eta_S^{(\xi)}]^L \sum_{I\supseteq L}[1 - \eta_S^{(\xi)}]^{I-L}\omega^{(I,\xi)} \tag{38}$$

$$p_S^{(\xi)}(\mathrm{x}, \ell) \ = \ \frac{\langle p_S(\cdot, \ell)\mathrm{f}(\mathrm{x}|\cdot, \ell), p^{(\xi)}(\cdot, \ell)\rangle}{\eta_S^{(\xi)}(\ell)} \tag{39}$$

$$\eta_S^{(\xi)}(\ell) \ = \ \left\langle p_S(\cdot, \ell), p^{(\xi)}(\cdot, \ell)\right\rangle \tag{40}$$

$[\cdot]_+$ stands for prediction, $\mathrm{f}(\mathrm{x}|\cdot, \ell)$ is the state transition function that we will propose in Section III-D. $\mathbb{B}$ is the space

[2] We briefly give necessary background on the GLMB in Subsections III-A to III-C. Readers are encouraged to refer to [32]–[34] and their references for detailed studies on GLMB, δ-GLMB, LMB Bayes RFS tracking filters.

[3] An RFS is a finite-set-valued random variable, whose number of points is random and the points are unordered and also random [45].

of new-born target labels. The set of new-born targets can be represented by an LMB RFS, where w_B is the probability of a birth hypothesis of new-born targets and p_B is the probability distribution of pitch states that belong to the birth targets as will be detailed in Section III-E. $p_S(\cdot, \ell)$ is the survival probability. The inclusion function, a generalization of the indicator function is defined as

$$1_Y(X) \triangleq \begin{cases} 1, \text{ if } X \subseteq Y \\ 0, \text{ otherwise.} \end{cases} \tag{41}$$

D. The Pitch Transition Model

A possible way of exploiting the temporal continuity for pitch tracking is using the hidden Markov model (HMM) [12], [20], [24]. While it is reasonable and useful to model the pitch transition as a Markov process, the HMM however, usually requires training the algorithms to obtain *a priori* knowledge of the state transition probabilities, which can be inconvenient and restricting in practice.

It is well-known that the pitch of a human speaker depends on the vocal tract, sub-glottal resonance and speech content. Because of the pronunciation, intonation and emotion, the fundamental frequency of a human speaker can vary over a continuous range [37]. This range is usually a limited subset of $[F0_{min}, F0_{max}]$, and naturally, the pitch of a human speaker tend to move toward its average level over time.

Therefore, assuming a Gaussian distribution of pitch states centered at the speaker's mean pitch value over time, and using the temporal continuity constraint, we propose to model the speaker pitch transition function $\mathrm{f}(\mathrm{x}_{\hat{f}}|\hat{f}, \ell)$ in (39) as an Ornstein-Uhlenbeck process [35], i.e.

$$\mathrm{x}_{\hat{f}} = \hat{f} + \alpha \cdot (\mu_{\hat{q}} - \hat{f}) \cdot t_{step} + \nu_{\sigma_{\hat{q}}}, \tag{42}$$

where $\hat{f} \in \hat{F}_0$ denotes a measurement of pitch (from the pitch estimator) at current time, and $\mathrm{x}_{\hat{f}}$ denotes the "predicted" pitch state at next time frame. Parameters $\mu_{\hat{q}} > 0$ and $\sigma_{\hat{q}} > 0$ are respectively the mean value and standard deviation of the pitch with index $\hat{q}$. The reversion rate $\alpha > 0$ specifies how fast the pitch return to its mean, and t_{step} ($t_{step} = n_{step}/f_s$) is the time step. We choose $\alpha = 0.1$ in this paper. $\nu_{\sigma_{\hat{q}}} \sim \mathcal{N}(0, \sigma_{\hat{q}})$ is the Gaussian distribution with mean value of 0 and standard deviation of $\sigma_{\hat{q}}$. Apparently (42) describes a Gaussian and Markov process [35] with long term mean of $\mu_{\hat{q}}$, hence is also called a mean-reverting process.

In (42), an arbitrary pitch measurement $\hat{f} \in [F0_{min}, F0_{max}]$ is mapped to an index $\hat{q}$ via:

$$\hat{q} = \arg\min_{q} |\hat{f} - \mu_q|, \ \mu_q \in \vec{\mu}, \tag{43}$$

where μ_q span the pitch range $[F0_{min}, F0_{max}]$ evenly with steps of $\mu_S > 0$, i.e.

$$\vec{\mu} = \{\mu_q \mid \mu_q = F0_{min} + \mu_S \cdot (q - 1/2), \ q = 1, ..., q_M\}, \tag{44}$$

where $q_M = \lfloor (F0_{max} - F0_{min})/\mu_S \rfloor$. This is reasonable since we have no *a priori* knowledge of the pitch level, sampling the pitch range with μ_q initializes the mean-reverting process.

Typically a greater σ_q corresponds to a greater μ_q. Thus with a first-order approximation of coefficient $\kappa_\mu \in (0, 1)$, we have

$$\sigma_q = \kappa_\mu \cdot \mu_q. \tag{45}$$

With a step size μ_S not too large, and the coefficient κ_μ not too small, reasonable sampling of pitch range can be obtained. We choose $\mu_S = 40$Hz and $\kappa_\mu = 0.1$ in this paper.

E. Measurement Driven Birth

The standard implementation of GLMB filter in (36) and (37) in Section III-C relies on *a priori* knowledge of target birth distributions, which restricts its applications in practice. Here we adapt the measurement-driven birth model that we presented in [36] for pitch tracking. It initiates the pitch states and existence probabilities of birth targets based on measurement data (pitch estimates) from previous time, hence adaptively tracks speaker pitches online. More details of measurement driven birth model for LMB and GLMB can be found in [33], [36] respectively.

In the GLMB update step, measurements $\hat{f} \in \hat{F}_0$ are associated with persistent tracks and the corresponding hypothesis probability $\omega^{(I,\xi,\theta)}(\hat{F}_0)$ in (31) as well as the probability density $p^{(\xi,\theta)}(\mathrm{x}, \ell|\hat{F}_0)$ in (32) are calculated. According to the corresponding hypothesis probability, each pitch measurement $\hat{f}$ initiates new-born targets at the next time step, with the new-born likelihood for each measurement $\hat{f} \in \hat{F}_0$ found by

$$r_N(\hat{f}) = 1 - \sum_{(I,\xi) \in \mathcal{F}(\mathbb{L}) \times \Xi} \ \sum_{\theta \in \Theta(I)} 1_{\hat{f}_\theta}(\hat{f}) \omega^{(I,\xi,\theta)}, \tag{46}$$

where the inclusion function indicates if the measurement $\hat{f}$ has been assigned to a target by any of the updated hypotheses. It can be seen from (46) that, a measurement that has been used in all hypotheses cannot initiate a new-born target ($r_N(\hat{f}) = 0$), while for measurements that have not been assigned to any of the targets, the new-born likelihood is 1.

For each measurement $\hat{f}$ that has non-zero new-born likelihood, a new birth of Bernoulli RFS is generated around the measurement, assuming a Gaussian distribution. Thus the probability distribution of the states $p_B(\mathrm{x}, \ell)$ in (37) for the measurement-driven birth model is given as,

$$p_B(\mathrm{x}, \ell; \hat{f}) = \sum_{i=1}^{M_b} \frac{1}{M_b} \delta_{\mathrm{x}_{\hat{f}}^{(i)}}(\mathrm{x}), \ \hat{f} \in \hat{F}_0 \tag{47}$$

$$\mathrm{x}_{\hat{f}}^{(i)} \sim \mathcal{N}\big(m_B(\hat{f}), P_B(\hat{f})\big), \ i = 1, ..., M_b \tag{48}$$

where M_b denotes the number of generated states for the birth target. $m_B(\hat{f}) = \mu_{q_B}$ where q_B is found from (43). $P_B(\hat{f}) = \sigma_{q_B}$ is a variance that specifies the distribution of states of the new-born target, and can be found from (45). Larger values of $P_B(\hat{f})$ result in higher error tolerance, while smaller values give better accuracy in general.

Thus the set of new-born targets is a labeled multi-Bernoulli RFS with the probability density given as [33]

$$\pi_B(\mathbf{X}_+) = \Delta(\mathbf{X}_+) w_B(\mathcal{L}(\mathbf{X}_+)) \, [p_B]^{\mathbf{X}_+}, \tag{49}$$

where the birth probability also required in (36) is

$$w_B(I)=\prod_{i\in\mathbb{B}}\left(1-r_B^{(i)}\right)\prod_{\ell\in I}\frac{1_{\mathbb{B}}(\ell)r_B^{(\ell)}}{1-r_B^{(\ell)}},\tag{50}$$

and the existence probability of the Bernoulli MDB at the next time that is initiated by a measurement $\hat{f}\in\hat{F}_0$ depends on its new-born likelihood obtained from current time:

$$r_B(\hat{f})=\min\left(r_{B_{\max}},\ \lambda_B\cdot\frac{r_N(\hat{f})}{\sum_{\zeta\in\hat{F}_0}r_N(\zeta)}\right),\tag{51}$$

where λ_B is the expected number of target birth at the next time, and $r_{B_{\max}}\in(0,1]$ is the maximum existence probability of a new-born target to ensure that the resulting $r_B(\hat{f})$ does not exceed 1 when λ_B is too large. We choose $M_b=1000$, $\lambda_B=0.3$ and $r_{B_{\max}}=0.15$ in this paper as in [36].

F. Adaptations for Pitch Tracks

The above implementation of the GLMB filter can produce the number of pitch targets and the estimates of target pitch with respective labels (identities) over time. However, when pitch tracks are well apart in time, they tend to be assigned with different target identities, even if they are from the same speaker. Moreover, for the single-pitch tracking, it may still produce two or more targets due to spurious errors. Thus we propose further adaptations here for these two cases.

1) Labeling Adaptation: In practice, it is common that a same speaker can have pauses during speech thus should be assigned with a same label.

Assuming that pitches of a close level belong to one target, we provide an adaptation for target labeling here. Once a new pitch target is confirmed, we compare its pitch estimate with all previously confirmed tracks. If the smallest difference is less than 20% of the mean value of a particular pitch track, we assign the label of that track to the new target and update the association map by marking the pause or unvoiced periods as miss detections.

2) Single Pitch Extraction: The single pitch extraction adaptation is proposed by selecting the pitch target with the highest accumulated probability from all hypotheses.

The pitch estimate is found with the label

$$\hat{\ell}=\arg\max_{\ell}w(\ell),\tag{52}$$

where directly from the definition in (30), the accumulated probability of all hypotheses containing label ℓ is found, i.e.

$$w(\ell)=\sum_{I\ni\ell}\omega^{(I,\xi,\theta)}(\hat{F}_0).\tag{53}$$

IV. Numerical Studies

This section demonstrates the performance of the proposed pitch estimator and tracker under various conditions. We first provide the pitch estimation and tracking results for the case of a single speaker, in presence of additive noises at various levels of signal to noise ratios (SNR). Then we also provide multi-pitch estimation and tracking results for concurrent speakers. The sound corpora used are from the CSTR database [46], [47], which include 50 English utterances from a male and a female speaker respectively and their laryngograph signals. The Keele database is also used for verification [48]. The noise signals used include the white Gaussian noise as well as those from the AURORA database [49], [50], which are composed of 8 types of noises from different environments.

A. Experimental Setup

All sound signals are resampled first at $f_s=16000$Hz. For the pitch tracker, the detection probability for the pitch estimator is modeled as $p_D(\mathrm{x},\cdot)\sim p_{D_{max}}\cdot\mathcal{N}(\mathrm{x};f_{mid},R^2)$, where $p_{D_{max}}=0.98$, $f_{mid}=\frac{1}{2}\cdot(F0_{min}+F0_{max})$ and $R=100$Hz. The measurement likelihood for the GLMB update is $g(\hat{f}|x)\sim\mathcal{N}(\hat{f};x,D)$, where $D=5$Hz. The survival probability is $p_S=0.8$, and the clutter rate is $\kappa=0.0001$. The single speaker adaptation is applied for the case of single speaker, while the labeling adaptation is used for the case of multi-pitch tracker for concurrent speakers.

B. Performance Metrics

We use the standard gross pitch error (GPE) for evaluating the accuracy of pitch estimates in voiced regions [12], [17].

$$\mathrm{GPE}\triangleq\frac{N_{err}}{N_v},\tag{54}$$

where N_{err} is the number of frames with pitch estimates that deviate from ground truth by more than 5%, and N_v denotes the total number of voiced frames as reported by both the ground truth and the estimation method.

The voicing decision error (VDE) metric is also used to evaluate the accuracy in deciding voiced/unvoiced frames, i.e.

$$\mathrm{VDE}\triangleq\frac{N_{ue}+N_{ve}}{N_f},\tag{55}$$

where N_{ue} is the number of frames that have pitch estimates but are unvoiced from ground truth, N_{ve} is the number of frames that have no pitch estimates but are actually voiced, and N_f is the total number of frames.[4]

For multi-pitch tracking, the GPE or VDE measures may not suffice. Thus we evaluate the performance also with the speaker identity error (SIE), i.e.

$$\mathrm{SIE}_i\triangleq\frac{E_{ij}}{N_{v_i}},\ i,j\ \in\ \{1,2,\cdots\},\tag{56}$$

where N_{v_i} denotes the number of voiced estimates for speaker i reported by both the ground truth and the pitch tracker, and E_{ij} denotes the number of pitch estimates that are assigned to speaker i, but actually belong to speaker j.

C. Single Speaker

Fig. 4 shows the speech signal and pitch estimation results for a male speaker with additive babble noise, using the RAPT (Robust Algorithm for Pitch Tracking) [2], YIN [4], PEFAC (Pitch Estimation Filter with Amplitude Compression) [11],

[4] Note that for estimators that produce a pitch estimate in every frame (e.g. PEFAC or YIN), $N_{ve}\equiv 0$, and the resulting VDE may be biased toward the ratio of the total number of unvoiced frames to N_f, if the respective voicing decision measure (e.g. voice probability or aperiodicity) is not used.

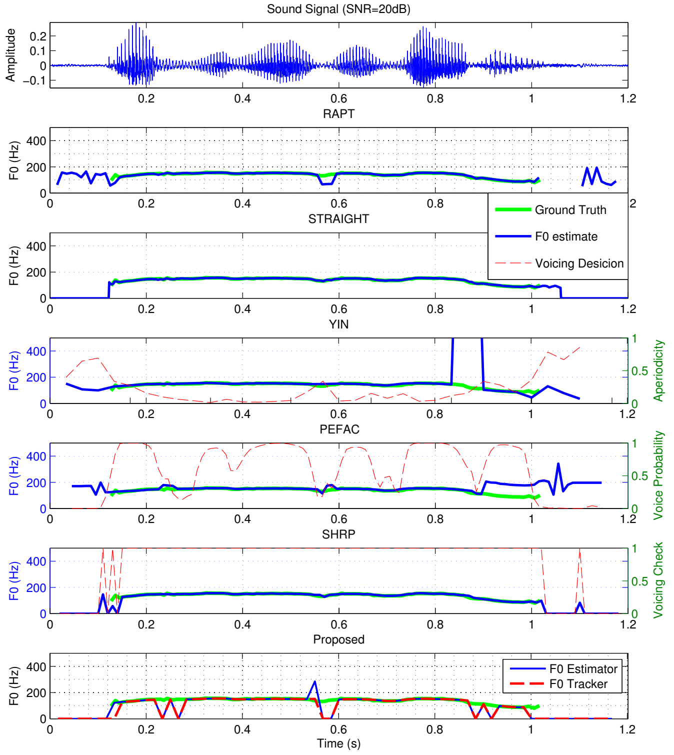

Fig. 4. From top to bottom: waveform of the speech signal with babble noise at SNR of 20dB (top panel), pitch ground truth of clean speech signal and pitch estimation results from the RAPT, STRAIGHT, YIN, PEFAC, SHRP and the proposed methods, respectively.

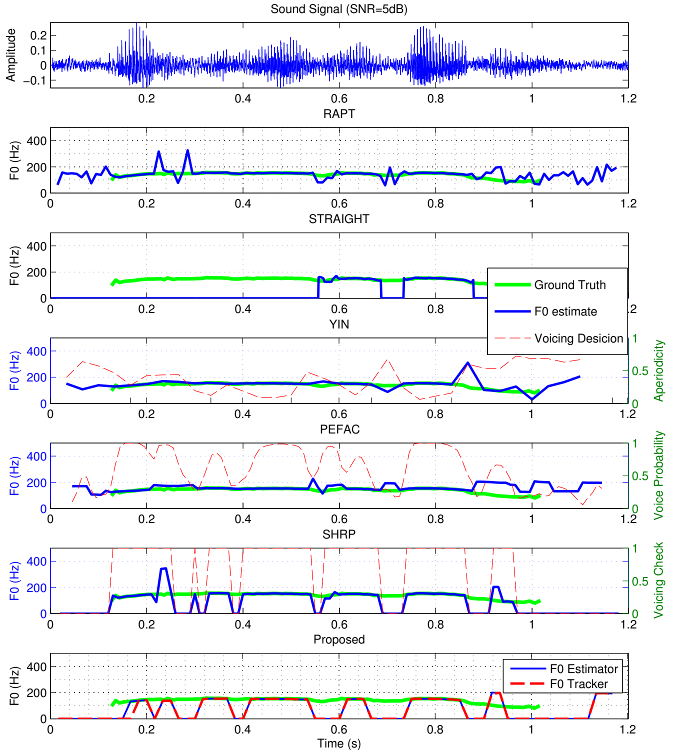

Fig. 5. From top to bottom: waveform of the speech signal with babble noise at SNR of 5dB (top panel), pitch ground truth of clean speech signal and pitch estimation results from the RAPT, STRAIGHT, YIN, PEFAC, SHRP and the proposed methods, respectively.

SHRP (Subharmonic-to-Harmonic Ratio based Pitch determination algorithm) [10], STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectrum) and our proposed methods, at SNR of 20dB. Pitch ground truth is plotted in green as reference for each method. This figure represents the cases when the noise is weak. All the methods can produce accurate pitch estimates during the voiced period for most of the time, compared to the ground truth. The RAPT method produces spurious estimates at about 0.56s, and also for the babble noise before about 0.1s and after about 1.1s. The STRAIGHT produces perfect estimates in this case, except the tail after about 1s. The YIN method uses the "Aperiodicity" measure as a voiced/unvoiced sound detector, however, a low aperiodicity can also correspond to erroneous estimates at around 0.85s. The PEFAC method provides "Voice Probability" for detection of voiced/unvoiced sounds. The SHRP method provides a binary "Voicing Check" measure. The "Aperiodicity", "Voice Probability" and "Voicing Check" can be used as the voiced/unvoiced activity detector (VAD) for respective methods, and are plotted in red. Both the PEFAC and SHRP methods are frequency domain methods, and their estimates are regarded reliable when the corresponding voicing decision measures are high (i.e. "Voice Probability" is close to 1 or the "Voicing Check" equals 1). However, we can also see that these voicing decision measures may also have outliers. For example, the PEFAC produces correct estimates at time of about 0.27s and 0.85s while its voicing decision measure is close to 0. The SHRP produces inaccurate estimates at about 0.15s and after 1s where its voicing decision measure is 1. The proposed pitch estimator and single pitch tracker produces comparatively reliable results. The pitch estimates from the proposed pitch estimator and pitch tracker are all close to ground truth. There are miss-detections (i.e. empty set of estimates, plotted as zeros for clarity) at about 0.25s, 0.55s and 0.9s, which correspond to the time segments when the voiced speech signal is weak. In this case, our proposed method produces no estimate for the time period dominated by weak (SNR=20dB) babble noise (i.e. before about 0.1s and after about 1.1s). Note also that there is a spurious estimate at about 0.55s, and the pitch tracker successfully filters it.[5] When there is no spurious estimate from the pitch estimator, the pitch estimates from the proposed pitch tracker almost overlap with those from the pitch estimator. The adaptive measurement driven birth model of the pitch tracker requires the initial measurements before confirming a new track, which can be seen at time of about 0.15s.

Fig. 5 shows the case when the additive noise is strong (SNR= 5dB). All existing methods produce spurious pitch estimates during the voiced period, compared with the ground truth. The RAPT produces considerable errors over time, which can be hard to suppress. The STRAIGHT produces correct pitch estimates between about 0.55s and 0.9s. The

[5] Note however that although the pitch tracker is useful at higher SNRs (e.g. SNR$\geq$ 0dB), it may not be able to improve the pitch estimation performance at very low SNRs (e.g. SNR$\leq$ -5dB) due to excessive spurious estimates and miss-detections.

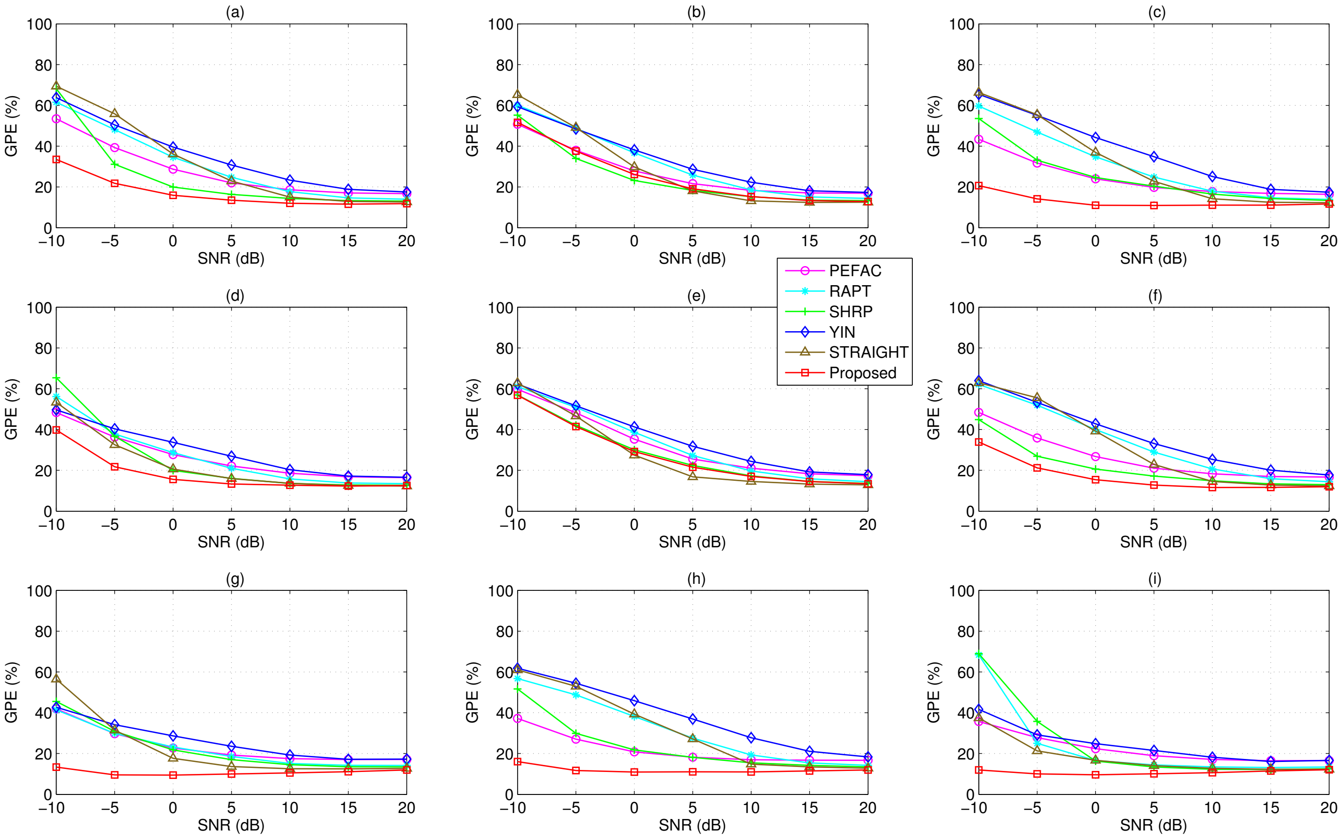

Fig. 6. Averaged GPE results for pitch estimation. Speech signals are from the CSTR and Keele corpora, while the additive noises are from the Aurora database and AWGN. Noise types are respectively (a) airport, (b) babble, (c) car, (d) exhibition, (e) restaurant, (f) street, (g) subway, (h) train, (i) AWGN.

other methods can take advantage of their corresponding voice decision measures. The YIN uses the aperiodicity as its voicing decision measure, i.e. the estimates can be removed when the aperiodicity is high. Comparing Fig. 4 and Fig. 5, in this paper, any estimate corresponding to an aperiodicity of greater than 0.5 is discarded. For PEFAC, we choose the estimates with a voice probability of no less than 0.5. For SHRP, all estimates corresponding to voice check of 1 are used. Our proposed pitch estimator and tracker can produce accurate pitch estimates during the voice period, although having more miss-detections compared with that of Fig. 4 due to the stronger babble noise. All estimates from our proposed methods are used for further quantitative comparison using the GPE metric. Note that by applying the VADs of respective reference methods, the GPE evaluates the accuracy of all selected pitch estimates, without taking into account those discarded ones. In general, all the five state-of-the-art methods (RAPT, YIN, PEFAC, STRAIGHT and SHRP) can provide reasonably reliable pitch estimates for high SNR sounds, but show different levels of performance degradation as the SNR drops. The spurious estimates can be suppressed to different extents, using their corresponding VADs. Overall, our proposed pitch estimator has produced reliable pitch estimates for voiced speech. For the sake of clarity in Fig. 4 and Fig. 5, when the estimate is an empty set, we plot the pitch value as a zero. Similar to the reference methods, discarded pitch estimates (empty sets of pitch estimates in our proposed methods) are not counted in the GPE measure.

In Fig. 6, we show the GPE results for all the methods using the CSTR and the Keele corpora with various types of noise and SNR levels. The GPEs are averaged over all sound files. It may be arguable as to the fair selection of parameter values for best performance of respective methods. Here the parameters for RAPT, YIN, STRAIGHT, PEFAC and SHRP all use the default values as provided in respective programs, and in particular, frame lengths are 10ms, 33.3ms, 80ms, 10ms and 10ms respectively. We choose a frame length of 33.3ms for our proposed pitch estimator, which is two periods of the minimum F0 frequency ($F0_{min} = 60$Hz). We can see from Fig. 6 that all methods degrade as the noise get stronger. However, the proposed pitch estimator outperforms the other state-of-the-art methods in most cases. The performance of the proposed method is worst at the babble noise or the restaurant noise, both of which are basically random mixtures of human speech signals. It is also interesting to notice that the STRAIGHT method performs better than most other methods at high SNRs. Moreover, compared with other noise types, the additive white Gaussian noise (AWGN) seems to cause least degradation to all these pitch estimators, except the outliers from the RAPT and SHRP at SNR≤ -5dB.

The proposed frequency coverage is verified in Fig. 7 where we test and check the GPE results for the male and female speakers of the CSTR corpus at various frequency coverage. Here the sound signals are the same as used in Fig. 6, but the GPE is an averaged result over all noise types. We can clearly see that despite the changes of SNR, the accuracy improves (the gross pitch error decreases) as η_C increases until $\eta_C = 1$, and the GPE is comparatively stable for $\eta_C \in [1, 1.5]$. We know that as η_C increases, the number of subbands also increases, thus requiring more computations. Hence as we have expected in Section II-C, $\eta_C = 1$ is chosen for good estimation accuracy and low computational load for our proposed methods.

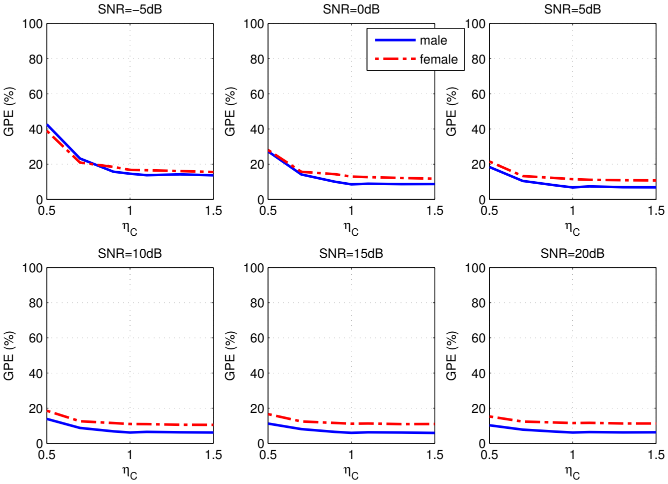

Fig. 7. GPE versus frequency coverage η_C using pitch estimation results (averaged over all noise types and sound files).

In Fig. 8 we show the VDE results. We can see that the VDE decreases as the SNR increases in general for all methods. The proposed method performs consistently in all the test cases. For high SNRs, the SHRP seems to perform the best. The RAPT seems to have difficulties with the "speech-like" noise types even at high SNRs (cf. Fig. 4 and Fig. 5).

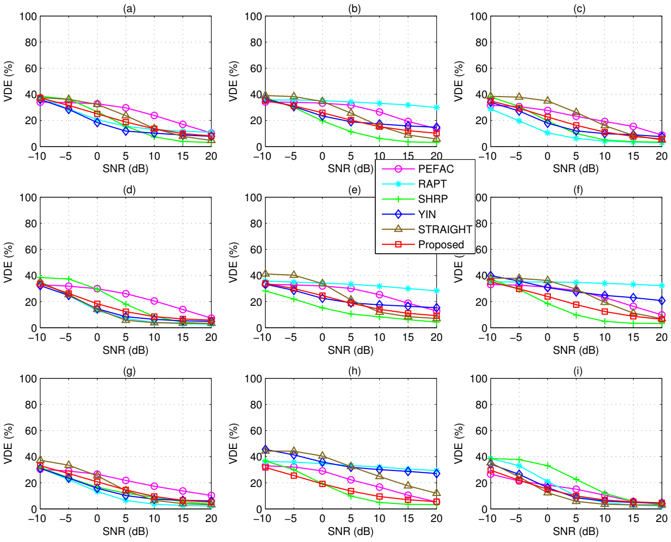

Fig. 8. VDE results using various methods for pitch estimation. Speech signals are from the CSTR database, while the additive noises are from the Aurora database and AWGN. Noise types are respectively (a) airport, (b) babble, (c) car, (d) exhibition, (e) restaurant, (f) street, (g) subway, (h) train, (i) AWGN.

D. Reverberation

Reverberation can also cause errors to most pitch estimation algorithms, because the reflections change the waveform and spectra of the sounds. In Fig. 9, we show an example of real recordings of the speech signals from the CSTR database in an office room with reverberation time of $T_{60} \approx 0.65$s. A loudspeaker is used to play the original sound corpora, and the electret omnidirectional microphone, preamp and sound card are used for recording. We can see that the reverberation creates long "tails" in the waveforms and spectra of the sound recording, especially during speech pauses, which is also obvious in the pitch estimation results using our proposed pitch estimator and tracker (see the bottom panel), compared with the ground truth for the clean speech signal. However, the values of pitch estimates are close to the ground truth over time. Fig. 9 also shows the pitch estimation results from the RAPT, STRAIGHT, YIN, PEFAC and SHRP methods respectively, using the reverberant recording. We can see that all the methods (except the SHRP) produces "tails" due to the reverberation. The RAPT has some spurious estimates at around 0.8s. Estimates of the STRAIGHT overlap well with ground truth, except for the miss-detections at about 0.8s and 1.2s. The YIN and PEFAC, considering also their corresponding voicing decision measures, produce accurate pitch estimates. The SHRP however, has considerable miss-detections, although all its voiced estimates are accurate.

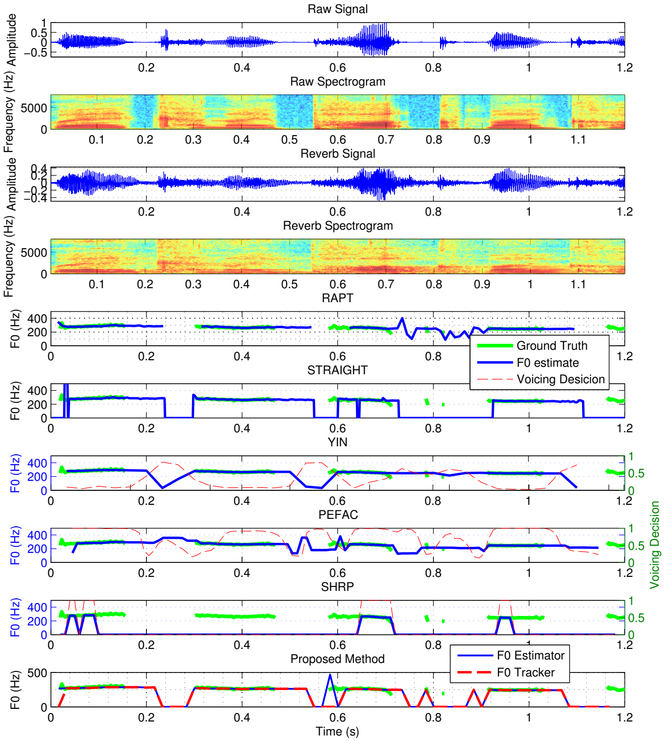

Fig. 9. Speech signals of a female speaker and its reverberant recording (top four panels), and pitch estimates from the RAPT, STRAIGHT, YIN, PEFAC, SHRP and the proposed methods respectively. No additive noise.

E. Multiple Speakers

We also test our proposed multi-pitch tracking method for concurrent speakers. Since the GLMB filter is capable of tracking states of a time-varying number of objects, the proposed pitch tracker is applicable to the scenario with more than two concurrent speakers, provided they are at different pitch levels. In this paper, we focus on the scenario of two speakers as it is more common. Fig. 10 shows pitch estimation results from our proposed multi-pitch tracker for the case when a female speaker and a male speaker (at different pitch levels) talk concurrently. The normalized and superimposed speech signals are from the CSTR corpus. Additive babble noise of various levels is included to test the reliability of the proposed method. We can see that at each SNR level, the proposed pitch tracker produces two separate pitch tracks and correctly assigns different labels (as shown in different colors) to the pitch estimates of the two respective speakers. There are unvoiced periods in the speech signals, but we can see that assigning the correct labels to the pitch estimates links the segments and forms a pitch track for each corresponding speaker. Moreover, the spurious pitch estimates from our proposed pitch estimator are filtered by the proposed pitch tracker since they do not form temporal continuity with their neighbouring pitch estimates. When the noise is weak, the majority of the pitches are detected and most of the pitch estimates are accurate compared to the ground truth. As the noise gets stronger, there are more miss-detections. There are also brief miss-detections due to competing sounds, especially when the noise is strong. However, most of the pitch estimates from the proposed pitch tracker are still close to ground truth.

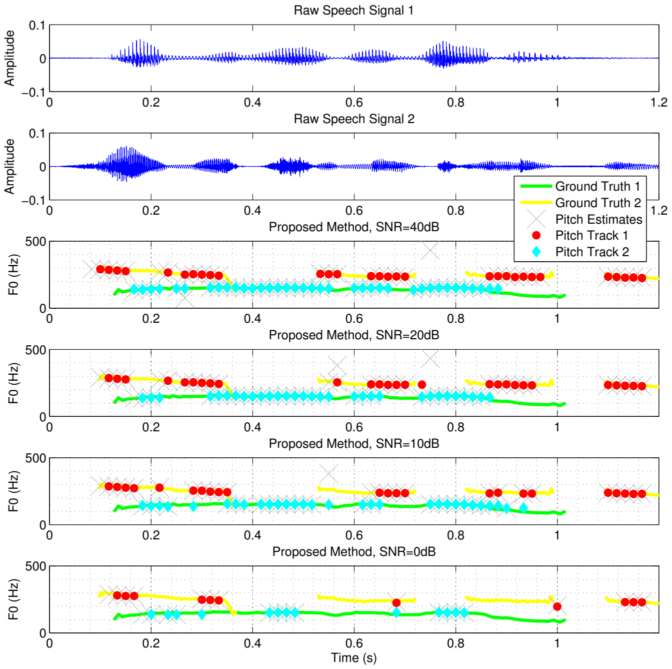

Fig. 10. Raw speech signals of a male speaker and a female speaker (top two panels), and pitch estimation and tracking results of the mixture of the two concurrent speakers, with additive babble noise of SNR=40, 20, 10 and 0dB (bottom four panels).

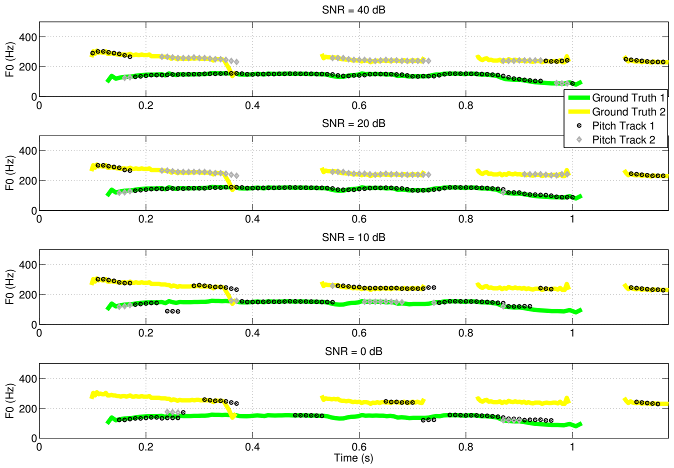

Fig. 11. Multi-pitch tracking results using Wu's method for the two concurrent speakers, with additive babble noise of SNR=40, 20, 10 and 0dB, respectively.

For comparison, Fig. 11 shows the multi-pitch tracking results of the same noisy speech mixtures at various SNRs, using Wu's method [20]. The pitch ground truth is plotted in each panel as reference. Wu's method produces two tracks of pitch estimates, i.e. Pitch Track 1 and Pitch Track 2, which are indicated with different colors. Here the available released C-code for Wu's method [51] is used as is, where the time step is 10ms and the time frame length is 16ms. Note that the results may not be the best possible from Wu's method, as ideally its parameters could be trained with the new database. We can see that at SNR=40dB, most of the pitches are extracted, but there are errors of pitch identities before 0.2s and after 0.9s. As the babble noise gets stronger, e.g. SNR=10dB and 0dB, there are more miss-detections and pitch identity errors, and almost all the identity of estimates of Pitch Track 2 are mistaken. We can also see in this case that the pitch identity errors of Wu's method do not always happen at the same time over different SNRs.

TABLE I
GPE of Wu's and the proposed multi-pitch tracking.

	Wu's		Proposed	
SNR (dB)	Speaker 1	Speaker 2	Speaker 1	Speaker 2
40	0.2273	0.1333	0.0667	0.1081
20	0.1705	0.1333	0.0385	0.0968
10	0.5625	0.9286	0	0.2333
0	0.6364	1	0.1	0.3077

Taking into account errors on the pitch labeling (identities) as well as the pitch accuracy, Table I gives the GPE results for the multi-pitch tracking using the proposed method and the Wu's method. We can see that the proposed method has less GPE compared with Wu's method over the range of SNRs, although in general the errors also tend to increase as the noise gets stronger. Table II shows the corresponding VDE results, which indicates that the proposed method has more voicing decision errors (mostly miss-detections) than Wu's method for the multi-pitch scenario (cf. Fig. 10 and Fig. 11).

Accurately filtering raw pitch estimates and correctly associating estimates with respective speakers are crucial features

TABLE II
VDE OF WU'S AND THE PROPOSED MULTI-PITCH TRACKING.

Methods	SNR = 40 (dB)	20	10	0
Wu's	0.0085	0.0254	0.0847	0.3305
Proposed	0.0429	0.1143	0.2000	0.5143

of the multi-pitch tracker. Thus we also measure the speaker identity errors of the pitch trackers as defined in (56), i.e. for each speaker, the ratio between the number of pitch estimates that actually belong to another speaker (see e.g. the pitch estimates before 0.15s at SNR= 40dB in Fig. 11), and its total number of pitch estimates. From Fig. 10 as expected, since the two speakers are at different pitch levels, there is zero SIE from the proposed method. However, from Fig. 11, there are considerable identity errors from Wu's method especially at low SNRs. The SIE results are provided in Table III. Overall, although having higher VDEs, the proposed method provides considerably better GPEs and SIEs in the studied scenario.

TABLE III
SIE OF WU'S AND THE PROPOSED MULTI-PITCH TRACKING.

	Wu's		Proposed	
SNR (dB)	Speaker 1	Speaker 2	Speaker 1	Speaker 2
40	0.1136	0.1111	0	0
20	0.0568	0.0889	0	0
10	0.4500	0.9286	0	0
0	0.2364	1	0	0

V. CONCLUSION

In this paper we propose a new pitch estimator inspired by CASA approaches and a novel pitch tracker that does not require training. The pitch estimator uses an auditory filterbank to decompose the speech mixture. The number of subbands and center frequencies of the filterbank are calculated according to our proposed frequency coverage metric for consistent and full frequency coverage without redundancy. For reliable and distinct pitch estimates, it encodes subband signals before the autocorrelation operation. To further suppress spurious errors and connect pitch tracks of respective speakers, the pitch tracker is proposed based on the GLMB framework, assuming temporal continuity of pitch. We propose a novel pitch transition model based on the Ornstein Uhlenbeck process, and use the measurement driven birth model for adaptive tracking. We also provide some necessary adaptations of GLMB, including the single pitch tracking as well as the labeling in presence of unvoiced periods or long pauses during speech. Not only for single pitch tracking, this training-free pitch tracker is also applicable for multi-pitch tracking as long as pitches of concurrent speakers are on different levels.

The proposed pitch estimator and tracker produce not only reliable pitch estimates but also pitch labels (speaker identities) for respective speakers. Evaluations using the CSTR, Keele and AURORA databases as well as real recordings in a reverberant room of $T_{60} \approx 0.65s$ have demonstrated the reliability of the proposed methods against various additive noises and also reverberation. Numerical comparisons with other baseline methods also validate the benefits of the proposed methods.

ACKNOWLEDGMENT

The author would like to acknowledge the contributions of the Australian Postgraduate Award and Australian Government Research Training Program Scholarship in supporting this research. The author is grateful to Dr. DAn Ellis for the sound database that he generously shared and the intriguing discussions on the binary file conversion, to Dr. T.C. Toh for his helpful mathematical insights and review comments, and to anonymous reviewers for constructive comments that help improve the quality of the paper.

APPENDIX A
EXPRESSION OF THE SUBBAND SIGNAL

From (2), the speech harmonic component is

$$\begin{aligned} s_q^{(\hbar)}(t) &= A_q^{(\hbar)}(t) \cdot \cos\left(\hbar \cdot \omega_q \cdot t + \phi_q^{(\hbar)}(t)\right) \\ =&\frac{1}{2} A_q^{(\hbar)}(t) \cdot [e^{\mathrm{i}[\hbar \cdot \omega_q \cdot t + \phi_q^{(\hbar)}(t)]} + e^{-\mathrm{i}[\hbar \cdot \omega_q \cdot t - \phi_q^{(\hbar)}(t)]}] \end{aligned} \tag{57}$$

where $\omega_q \triangleq 2\pi f_q$.

Using linear-phase filters, e.g. the gammatone filter, from (4) we have

$$\begin{aligned} g^{(b)}(t) &= \tilde{g}^{(b)}(t) \cdot \cos(2\pi f_C^{(b)} t) \\ &= \frac{1}{2} \cdot \tilde{g}^{(b)}(t) \cdot (e^{\mathrm{i}2\pi f_C^{(b)} t} + e^{-\mathrm{i}2\pi f_C^{(b)} t}). \end{aligned} \tag{58}$$

From (2), (4) and (7), when $\hbar \cdot \omega_q \approx 2\pi f_C^{(b)}$, the subband signal is given as follows:

$$\begin{aligned} & x^{(b)}(t) \approx [s_q^{(\hbar)}(t - t_{d_q}) \cdot \mathrm{h}_{qi}(t_{d_q})] * g^{(b)}(t) \\ & = \mathrm{h}_q(t_{d_q}) \cdot [\frac{1}{2} \cdot \tilde{g}^{(b)}(t) \cdot (e^{\mathrm{i}2\pi f_C^{(b)} t} + e^{-\mathrm{i}2\pi f_C^{(b)} t})] \\ & \quad * [\frac{1}{2} A_q^{(\hbar)}(t - t_{d_q}) \\ & \cdot [e^{\mathrm{i}[\hbar \cdot \omega_q \cdot (t - t_{d_q}) + \phi_q^{(\hbar)}(t - t_{d_q})]} + e^{-\mathrm{i}[\hbar \cdot \omega_q \cdot (t - t_{d_q}) - \phi_q^{(\hbar)}(t - t_{d_q})]}] \\ & \approx \frac{1}{4} \mathrm{h}_q(t_{d_q}) \cdot [A_q^{(\hbar)}(t - t_{d_q}) * \tilde{g}^{(b)}(t)] \\ & \cdot [e^{\mathrm{i}[\hbar \cdot \omega_q \cdot (t - t_{d_q}) + \phi_q^{(\hbar)}(t - t_{d_q})]} + e^{-\mathrm{i}[\hbar \cdot \omega_q \cdot (t - t_{d_q}) - \phi_q^{(\hbar)}(t - t_{d_q})]}]^{\dagger} \\ & = \frac{1}{2} \mathrm{h}_q(t_{d_q}) \cdot [A_q^{(\hbar)}(t - t_{d_q}) * \tilde{g}^{(b)}(t)] \cdot \\ & \qquad \cos(\hbar \omega_q (t - t_{d_q}) + \phi_q^{(\hbar)}(t - t_{d_q})) \\ & = \tilde{S}_q^{(b)}(t) \cdot \cos(\tilde{\phi}_q^{(b)}(t)),\ t \geq t_q, \end{aligned} \tag{59}$$

where

$$\tilde{S}_q^{(b)}(t) \triangleq \frac{1}{2} \cdot \mathrm{h}_q(t_{d_q}) \cdot A_q^{(\hbar)}(t - t_{d_q}) * \tilde{g}^{(b)}(t), \tag{60}$$

and

$$\tilde{\phi}_q^{(b)}(t) = 2\pi \hbar f_q (t - t_{d_q}) + \phi_q^{(\hbar)}(t - t_{d_q}). \tag{61}$$

■

[5]† Considering frequency domain meanings of convolution and complex exponentials for the Fourier transform and inverse Fourier transform.

Appendix B
Derivation of the ERBS Expression

Proof. From (12),

$$\begin{aligned}\Upsilon(f) &\triangleq \int \frac{1}{\upsilon(f)}df = \int \frac{1}{D + E \cdot f}df \\ =&\frac{1}{E} \cdot \ln(D + E \cdot f) + \text{Constant} \\ =&\frac{1}{E \cdot \lg e} \lg(D + E \cdot f) + \text{Constant} \\ =&\frac{1}{E \cdot \lg e}[\lg(1 + \frac{E}{D} \cdot f) + \lg D] + \text{Constant},\end{aligned} \tag{62}$$

which using the boundary condition that $\Upsilon(0) = 0$, leads to $\Upsilon(f) = \frac{1}{E \cdot \lg e}[\lg(1 + \frac{E}{D} \cdot f)]$, hence (12). ■

Appendix C
Frequency Coverage and Number of Subbands

Proof. From (13), (14) and (15), the frequency coverage is

$$\begin{aligned}\eta_C^{(b)} &\triangleq \frac{1}{2} \cdot \frac{f_B^{(b+1)} + f_B^{(b)}}{f_C^{(b+1)} - f_C^{(b)}} \\ &= K_\vartheta \cdot \frac{D + \frac{E}{2} \cdot (f_C^{(b+1)} + f_C^{(b)})}{f_C^{(b+1)} - f_C^{(b)}} \\ &= \frac{E{\cdot}K_\vartheta}{2} \cdot \Bigg[\Big((1 + D'f_{min})^{(N_b - b - 1)}{\cdot}(1 + D'f_{max})^{(b)}\Big)^{\frac{1}{N_b - 1}} + \\ &\quad \Big((1 + D'f_{min})^{(N_b - b)} \cdot (1 + D'f_{max})^{(b-1)}\Big)^{\frac{1}{N_b - 1}}\Bigg]\Bigg/ \\ &\quad \Bigg[\Big((1 + D'f_{min})^{(N_b - b - 1)} \cdot (1 + D'f_{max})^{(b)}\Big)^{\frac{1}{N_b - 1}} - \\ &\quad \Big((1 + D'f_{min})^{(N_b - b)} \cdot (1 + D'f_{max})^{(b-1)}\Big)^{\frac{1}{N_b - 1}}\Bigg] \\ &= \frac{E \cdot K_\vartheta}{2} \cdot \Big[(1 + D'f_{max})^{\frac{1}{N_b - 1}} + (1 + D'f_{min})^{\frac{1}{N_b - 1}}\Big]\Big/ \\ &\quad \Big[(1 + D'f_{max})^{\frac{1}{N_b - 1}} - (1 + D'f_{min})^{\frac{1}{N_b - 1}}\Big] \\ &= \frac{E \cdot K_\vartheta}{2} \cdot \frac{(\frac{D + E \cdot f_{max}}{D + E \cdot f_{min}})^{\frac{1}{N_b - 1}} + 1}{(\frac{D + E \cdot f_{max}}{D + E \cdot f_{min}})^{\frac{1}{N_b - 1}} - 1},\end{aligned} \tag{63}$$

where K_ϑ for the gammatone filter is a constant for a given filter order ϑ [41],

$$K_\vartheta = 2\sqrt{2^{1/\vartheta} - 1} \cdot \Big[\frac{\pi(2\vartheta - 2)!2^{-(2\vartheta - 2)}}{(\vartheta - 1)!^2}\Big]^{-1}. \tag{64}$$

Thus the number of subbands N_b as given in (16) can be directly obtained from (63). ■

References

[1] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[2] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

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

[4] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[5] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, 2013.

[6] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic surn spectrum, and a maximum likelihood estimate," in *Symposium on Computer Processing in Communication, ed.*, vol. 19. University of Broodlyn Press, New York, 1970, pp. 779–797.

[7] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[8] H. Kawahara, H. Katayose, A. d. Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[9] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[10] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–333.

[11] S. Gonzalez and M. Brookes, "Pefac-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[12] D. Wang, C. Yu, and J. H. Hansen, "Robust harmonic features for classification-based pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 952–964, 2017.

[13] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1635–1644, 2007.

[14] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* IEEE, 2008, pp. 101–104.

[15] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "A robust and computationally efficient subspace-based fundamental frequency estimator," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 3, pp. 487–497, 2010.

[16] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 76–87, 2004.

[17] W. Chu and A. Alwan, "Safe: A statistical approach to f0 estimation under clean and noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 933–944, 2012.

[18] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[19] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE transactions on speech and audio processing*, vol. 8, no. 6, pp. 708–716, 2000.

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

[21] J. O. Smith III and J. S. Abel, "Bark and erb bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[22] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[23] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, no. 3, pp. 191–207, 1997.

[24] B. S. Lee and D. P. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Proc. INTERSPEECH, Portland, OR, USA*, Sep, 2012.

[25] M. K. Reddy and K. S. Rao, "Robust pitch extraction method for the hmm-based speech synthesis system," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1133–1137, 2017.

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

[27] D. Wang, P. C. Loizou, and J. H. Hansen, "F0 estimation in noisy speech based on long-term harmonic feature analysis combined with neural network classification," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[28] J. Zhang, J. Tang, and L.-R. Dai, "Rnn-blstm based multi-pitch estimation." in *INTERSPEECH*, 2016, pp. 1785–1789.

[29] Y. Liu and D. Wang, "Speaker-dependent multipitch tracking using deep neural networks," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 710–721, 2017.

[30] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[31] S. Lin, "A new frequency coverage metric and a new subband encoding model, with an application in pitch estimation," *Proc. Interspeech 2018*, pp. 2147–2151, 2018.

[32] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3460–3475, 2013.

[33] S. Reuter, B.-T. Vo, B.-N. Vo, and K. Dietmayer, "The labeled multi-bernoulli filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3246–3260, 2014.

[34] B.-N. Vo, B.-T. Vo, and D. Phung, "Labeled random finite sets and the bayes multi-target tracking filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6554–6567, 2014.

[35] C. W. Gardiner *et al.*, *Handbook of stochastic methods*. Springer Berlin, 1985, vol. 3.

[36] S. Lin, B. T. Vo, and S. E. Nordholm, "Measurement driven birth model for the generalized labeled multi-bernoulli filter," in *2016 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2016, pp. 94–99.

[37] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[38] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE transactions on signal processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[39] S. Lin, "Reverberation-robust localization of speakers using distinct speech onsets and multichannel cross correlations," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 11, pp. 2098–2111, 2018.

[40] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[41] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[42] F. Nolan, "Intonational equivalence: an experimental evaluation of pitch scales," in *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, vol. 39, 2003.

[43] R. Lyon, "A computational model of binaural localization and separation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8. IEEE, 1983, pp. 1148–1151.

[44] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702–711, 1986.

[45] R. P. Mahler, *Statistical multisource-multitarget information fusion*. Artech House, Inc., 2007.

[46] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching." *Proc. Eurospeech*, pp. 1003–1006, 1993.

[47] Available at http://www.cstr.ed.ac.uk/research/projects/fda/.

[48] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[49] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[50] Available at https://www.ee.columbia.edu/∼dpwe/sounds/noise/.

[51] Available at http://web.cse.ohio-state.edu/pnl/shareware/wu-tsap03/.