Stochastic-Deterministic Signal Modelling for the Tracking of Pitch in Noise and Speech Mixtures using Factorial HMMs

Matthew McCallum and Bernard Guillemin

Department of Electrical and Computer Engineering
The University of Auckland, Auckland, New Zealand

Abstract

Obtaining estimates of the fundamental frequencies associated with either noise or speech in noise/speech mixtures can be important in speech enhancement. Accurate simultaneous estimation of these can result in both an improved subjective quality as well as a higher signal to noise ratio (SNR) of the resulting speech. It is crucial with such an algorithm that each periodic component be reliably identified as either noise or speech. Further, the algorithm needs to be robust to changing SNR of the noisy speech arising from a range of environmental conditions. In this paper a multipitch tracking algorithm is proposed based on a stochastic-deterministic (SD) signal model in the complex short-time Fourier transform (STFT) framework, using a factorial hidden Markov model (FHMM). Unlike previous multipitch tracking algorithms based on FHMMs, the proposed algorithm performs well even when the levels of noise and speech differ significantly from those of the training data. This robustness is attributed in part to the flexible SD model employed. With this model, a priori information of noise and speech used to identify and track non-stationary periodic components is based primarily on their spectral envelope, not their absolute amplitude.
Index Terms: Multipitch tracking, hidden Markov model, Gaussian mixture model, speech enhancement.

1. Introduction

Estimating the fundamental frequency of audio signals is a well known problem in the research of speech signal processing. Solving this estimation problem well has a broad range of applications including speech coding [1], speaker identification [2], single channel source separation [3, 4] and speech enhancement [5]. For the case where the analysed signal contains a mixture of fundamental frequencies (potentially corrupted by broadband noise), the problem, known as multipitch tracking, is challenging and has attracted much attention over the last decade [6, 7, 8, 9]. Relatively little research has focussed on the design of multipitch tracking algorithms specifically for the application of speech enhancement. In this paper, it is discussed how many recently developed algorithms fail to meet the requirements of such an application, and a new algorithm is derived that better meets these requirements when compared to a recently developed and closely related algorithm.

A range of speech enhancement algorithms have been developed that assume knowledge of the fundamental frequency, f_0, of speech [10, 11, 4, 12] , and of noise [5]. Clearly there is potential for periodic components of both speech and noise to simultaneously exist, in which case the correct f_0 (related to noise, speech or both acoustic sources, depending on the algorithm) must be labelled reliably. In the case that this labelling fails (i.e., the f_0 of speech is labelled to be that of noise, or vice versa), the consequences can be significant. For example, this misidentification in the case of [5], will result in the removal of speech from the noisy speech signal, which is highly undesirable.

This research is supported by Capability funding from the MBIE Science and Innovation Group.

Considering the range of environmental conditions under which the aforementioned speech enhancement algorithms may need to operate (if applied in a mobile communication device for example), then the multipitch tracking algorithm used to track and label f_0 must also be capable of performing irrespective of any environmental variation. Perhaps the most common and expected of these variations is that of SNR. If the multipitch tracking algorithm's performance is notably compromised by such a variation, then the range of applications where this algorithm is useful is severely limited.

With the necessary features of noise/speech source labelling and robustness to varying SNR, the design of a multipitch tracking algorithm is somewhat challenging. Algorithms have been proposed that are robust to varying SNR [6], and able to label sources [8]. However, both of these features rarely exist in any single algorithm. The specificity of [8] to the data on which the algorithm is trained provides an excellent source discrimination capability, but results in a heavy dependence on the environmental conditions under which the algorithm is trained (including the amplitude/power of training data). In contrast the design of [6] is based on some fundamental auditory properties of periodic signals which makes it reliable under a range of environmental conditions, but a lack of ability to incorporate source specific data into the algorithm means it is unable to inherently discriminate between noise and speech f_0.

In this paper an algorithm is derived that makes use of an FHMM, which has recently been identified as a natural framework for the statistical analysis involved in multipitch tracking [13, 8]. Unlike these earlier works which model the log-magnitude spectrum directly, and hence are heavily dependent on the magnitude of noise/speech, here a STFT SD speech model is employed [14, 15, 12]. This allows a direct modelling of the deterministic components that are responsible for periodicity in speech (impartial to noise/speech magnitude or equivalently SNR). Furthermore, this decomposition results in the use of an exact signal interaction model under the assumptions made, unlike other approximations used to date [13, 8].

2. SD signal modelling

Both Wohlmayr et al. [8] and Bach and Jordan [13] apply Gaussian models to magnitudes and/or log-magnitudes of the STFT. While this highlights the mean value of such features well, it is known that these variables are not in fact Gaussian distributed [16, 17]. In this paper the complex valued STFT is considered. When the signal contains a set of harmonically related deterministic components, each of the variables at the output of the STFT is modelled well by a non-zero-mean complex Gaussian distribution [18, 5, 12].

Specifically, in this paper mixtures of signals of the follow-

ing form are considered,

$$x_s[n] = \sum_{l=1}^{L_s} r_{l,s}[n]\cos(2\pi l f_{0,s}[n]n + \phi_{l,s}[n]) + u_s[n] = b_s[n] + u_s[n]. \quad (1)$$

Here subscript s indexes each source in a signal mixture. The signal $b_s[n]$ represents a deterministic component of $x_s[n]$, $u_s[n]$ represents a zero-mean stochastic process and L_s represents the number of harmonics in source s, indexed by l. This signal model covers many typical models for speech signals such as [19, 20, 21, 11], and a wide range of noise sources [5].

The problem this paper endeavours to solve is, given the signal mixture, $y[n] = \sum_{s=1}^{S} x_s[n]$, estimate $f_{0,s}[n]$. This paper considers the case $S = 2$ only which is very common in recent multipitch tracking literature [6, 13, 22, 8]. In terms of the STFT framework, the following signal is considered,[1]

$$Y[k,m] = \sum_{n=0}^{N-1} y[n+m\tau]w[n]\exp\left\{\frac{-2\pi ikn}{K}\right\}$$

at each frequency domain index $0 \leq k < K$, and each window frame index m. Here, $w[n]$ represents the time domain windowing function, N is the windowing length, K is the DFT length, and τ represents the increment in time index between each successive frame. It is practical to estimate $f_{0,s}[m\tau]$ for only integer values of m, provided that they change slowly enough to be considered constant for windowing length N. Likewise, a similar restriction applies to the variables $r_{l,s}$ and $\phi_{l,s}$.[2]

The statistical characterisation of signals of the form in (1) is well known in the frequency domain [18], and is reflected within their STFT representation. That is, for a given STFT frame, $X_s[k]$ is described by $X_s[k] = B_s[k] + U_s[k]$, where,

$$B_s[k] = \left(\sum_{l=1}^{L_s} r_{l,s}^2\right)^{\frac{1}{2}} \cdot \sum_{l=1}^{L_s} \frac{e^{\frac{a_{l,s}}{2}}}{2}\left\{e^{i\phi_{l,s}}W_{f_{0,s}}[k] + e^{-i\phi_{l,s}}W_{1-f_{0,s}}[k]\right\} \quad (2)$$

and $W_{f_{0,s}}[k] = \sum_{n=0}^{N-1} w[n]\exp\{2\pi in(f_{0,s} - k/K)\}$ is the discrete Fourier transform (DFT) of $w[n]$ modulated by a complex exponential. The normalised log magnitude variables, $a_{l,s} = \log\left\{r_{l,s}^2 / \sum_{l=1}^{L_s} r_{l,s}^2\right\}$ will be considered for the purposes of noise/speech labelling in this paper because they are independent from the total energy of the deterministic signal components for a given source.

Based on the discussion in [18, 19], $U_s[k]$ may be considered independent for all k, and for each k this signal may be characterised by a zero-mean complex Gaussian distribution. If $B_s[k]$ is considered a deterministic quantity as in [18, 5], then here $X_s[k]$ is distributed according to,

$$p(X[k]|B_s[k]) = \frac{1}{\pi\lambda_s[k]}\exp\left\{\frac{-|X_s[k] - B_s[k]|^2}{\lambda_s[k]}\right\}$$

where $p(\cdot)$ denotes a probability density function (pdf) and $\lambda_s[k]$ is defined as $E\left\{|X_s[k] - B_s[k]|^2\right\}$. It is then trivial to derive the pdf of the signal mixture for each k, given harmonic amplitudes and phases ($r_{l,s}$ and $\phi_{l,s}$) and frequency $f_{0,s}$. In this paper the notation $[z_s]_{s=c}^{d}$ will be used to represent a vector $[z_c, z_{c+1}, \ldots, z_d]^T$. Hence, we define the vectors $\mathbf{a}_s = [a_{l,s}]_{l=1}^{L_s}$ and $\boldsymbol{\phi}_s = [\phi_{l,s}]_{l=1}^{L_s}$.[3] For the purposes of notational brevity, the harmonic structure variable set $(\mathbf{a}_1, \mathbf{a}_2, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ will be henceforth denoted $\mathbf{H}$ where appropriate, and the frequency pair $f_{0,1}, f_{0,2}$ will be denoted as f_0. With this notation, the aforementioned pdf is given by,

$$p(Y[k]|\mathbf{H}, f_0) = \frac{1}{\pi\lambda[k]}\exp\left\{\frac{-|Y[k] - B[k]|^2}{\lambda[k]}\right\},$$

where $B[k] = B_1[k] + B_2[k]$ and $\lambda[k] = \lambda_1[k] + \lambda_2[k]$. Note that for a given STFT framework (i.e. given $w[n]$, τ, N and K), the parameter $B_s[k]$ is entirely specified by $f_{0,s}$, $\mathbf{a}_s$ and $\boldsymbol{\phi}_s$. Due to the independence of the variables $Y[k]$ for all k, under the assumed model, then the joint pdf of all observations for a given STFT frame is,

$$p(\mathbf{Y}|\mathbf{H}, f_0) = \frac{1}{\pi^{K/2}\det(\boldsymbol{\lambda})^{1/2}}\exp\left\{|\mathbf{Y} - \mathbf{B}|^T \boldsymbol{\lambda}^{-1}|\mathbf{Y} - \mathbf{B}|\right\} \quad (3)$$

where $\mathbf{Y} = [Y[k]]_{k=1}^{K}$, $\mathbf{B} = [B[k]]_{k=1}^{K}$ and $\boldsymbol{\lambda} = \mathrm{diag}\left([\lambda[k]]_{k=1}^{K}\right)$.

In a typical FHMM pitch tracking framework, (3) may be considered the observation model (i.e., given $\mathbf{H}$). It may be noted here that this observation model was derived based upon the exact additive interaction of the complex STFT of two sources. This is unlike previous FHMM based multipitch tracking methods that rely upon source interaction approximations such as additive STFT magnitude models [13, 8], or the mixture maximisation model [23, 8]. In addition, any signal noise is explicitly characterised by $\boldsymbol{\lambda}$. Explicit noise consideration is important for multipitch tracking algorithms to remain robust to different environmental noise conditions.

In the following section it will be seen that the observations $\mathbf{a}_s$ combined with the transitional probability mass function (pmf) $P(f_{0,s}^m | f_{0,s}^{m-1})$, described in the following section, allow the algorithm to differentiate between different sources based on both the spectral envelope (due to $\mathbf{a}_s$) and the time varying behaviour (due to $P(f_{0,s}^m | f_{0,s}^{m-1})$) of a particular source for which it is trained. The definition of the deterministic component (i.e., in (2)), is somewhat arbitrary in that it may be defined to be any feature of the complex STFT that is predictable from measurements over any time interval. $\mathbf{a}_s$ is sufficient for the purposes of this paper, although further research into the use of the SD model for multipitch tracking may consider a range of features. Hence the implementation of the SD model is rather flexible, yet, regardless of the features used the fundamental algorithm and approach described in this paper remains the same.

3. Pitch tracking in the FHMM framework with the SD signal model

FHMMs consider the tracking of multiple independent Markov processes that each evolve independently, when the available information to do so may be considered a joint result of the combined states of all Markov processes. In the multipitch tracking problem $f_{0,s}^m$ takes on a set of discrete frequency values. These variables may be considered the hidden variables in the FHMM with the independent prior probability mass function (pmf) $P(f_{0,s}^0)$, and independent transitional pmf $P(f_{0,s}^m | f_{0,s}^{m-1})$. The observation pdf, $p(\mathbf{Y}, \mathbf{H}|f_0) = p(\mathbf{Y}|\mathbf{H}, f_0)\, p(\mathbf{H}|f_0)$, that completes the description of the FHMM is described later in this section.

[1] In this paper an upper case letter represents the STFT of the corresponding time-domain signal denoted in lower-case.

[2] For notational simplicity, the explicit dependency of these variables on frame number m will be indicated in superscript. Futhermore, the dependence on m will be dropped from the notation altogether where it is unnecessary. For example, $f_{0,s}[m\tau] = f_{0,s}^m = f_{0,s}$.

[3] Here, bold symbols represent vectors and $[\cdot]^T$ denotes the vector transpose.

It is proposed here to augment the direct complex STFT observations $\mathbf{Y}^m$, with the harmonic structure observations $\hat{\mathbf{a}}_s^m, \hat{\boldsymbol{\phi}}_s^m$ which may be considered online estimates of $\mathbf{a}_s^m, \boldsymbol{\phi}_s^m$.This is necessary not only to evaluate (3), but these estimates are also useful features in assigning pitch tracks to their respective sources as will be described later in this section. Given the set of observations $\mathbf{Y}^m, \hat{\mathbf{a}}_s^m, \hat{\boldsymbol{\phi}}_s^m$, the complete FHMM may be described as follows. If it is denoted $\{\varsigma\} = \bigcup_{m=1}^{M} \varsigma^m$, for an arbitrary sequence ς^m. Then the joint pdf of all observations and frequency states is, [4]

$$p\left(\{\mathbf{Y}\}, \left\{\hat{\mathbf{H}}\right\}, \{f_0\}\right) = p\left(\{\mathbf{Y}\}, \left\{\hat{\mathbf{H}}\right\} | \{f_0\}\right) p\left(\{f_0\}\right)$$

$$= P\left(f_{0,1}^0\right) P\left(f_{0,2}^0\right) \prod_{m=2}^{M} p\left(\mathbf{Y}^m \left| \hat{\mathbf{H}}^m, f_0^m\right.\right) \cdot p\left(\hat{\mathbf{H}}^m | f_0^m\right) P\left(f_{0,1}^m | f_{0,1}^{m-1}\right) P\left(f_{0,2}^m | f_{0,2}^{m-1}\right). \quad (4)$$

Assuming that the harmonic structure variables composing $\hat{\mathbf{H}}^m$ are independent of each other, then $p\left(\hat{\mathbf{H}}^m | f_0^m\right)$ is the product $p\left(\hat{\mathbf{a}}_1^m | f_{0,1}^m\right) p\left(\hat{\mathbf{a}}_2^m | f_{0,2}^m\right) p\left(\hat{\boldsymbol{\phi}}_1^m | f_{0,1}^m\right) p\left(\hat{\boldsymbol{\phi}}_2^m | f_{0,2}^m\right)$. The multipitch tracking problem may then be solved by finding,

$$\{f_{est,1}, f_{est,2}\} = \arg\max_{\{f_0\}} p\left(\{\mathbf{Y}\}, \left\{\hat{\mathbf{H}}\right\}, \{f_0\}\right). \quad (5)$$

For an equivalent HMM the exact solution may be computed using the Viterbi algorithm, or alternatively using the junction tree algorithm [24]. To allow operation in real-time, the proposed algorithm considers the solution at each STFT frame over current and past data only, i.e., the solution to (5) over the sequence $0 \leq m \leq \mu$, where μ is the current frame.

For the observation model in (3) to be useful within an FHMM framework, the statistical properties characterising the variables $\mathbf{a}_s, \boldsymbol{\phi}_s$ upon which the observation model is dependent, must be specified. Previously, Wohlmayr overcame this requirement by modelling the magnitude and log-magnitude of $Y[k]$ directly and independently for each frequency state $f_{0,s}$, with the use of GMMs [8]. These magnitude or log-magnitude GMMs inherit the environmental characteristics of the training data (acoustic environment, equalisation, source levels etc.) and are not robust to changes in these.

Bach and Jordan [13] modelled the harmonic amplitudes $\mathbf{r}_s = [r_s]_{s=1}^{L_s}$ as a smooth Gaussian process on the line $[0, K/2]$. However in the application of speech enhancement, the assumed smooth set of harmonic amplitudes that is characteristic of speech may not accurately represent acoustic noise sources such as chainsaws or outboard boat engines [5].

To avoid these restrictions on source type and environment. The distribution of $\mathbf{a}_s$ is modelled (as opposed to $\mathbf{r}_s$) with the use of a set of GMMs and no smooth restriction, i.e.,

$$p(\mathbf{a}_s^m | f_{0,s}^m) = \sum_{q=1}^{Q} \alpha_{q,s} \mathcal{N}(\boldsymbol{\pi}_{q,s}, \boldsymbol{\Sigma}_{q,s}) \quad (6)$$

here, $\mathcal{N}(\boldsymbol{\pi}_{q,s}, \boldsymbol{\Sigma}_{q,s})$ represents an L_s dimensional Gaussian distribution function with mean $\boldsymbol{\pi}_{q,s}$ and covariance $\boldsymbol{\Sigma}_{q,s}$. Here $\boldsymbol{\Sigma}_{q,s}$ is restricted to be diagonal for all q. The GMM parameters of (6) may be trained using the expectation maximisation method [25]. The maximum description length (MDL) criteria is used to find an optimal Q as specified in [8].

The importance of signal phase (i.e., that of $\boldsymbol{\phi}_s$), is much debated in STFT audio processing literature [26]. Whilst the explicit characterisation of signal phase may still yield an improvement in performance for multipitch tracking algorithms, here the phase is considered to have an uninformative prior, i.e., it is assumed $\boldsymbol{\phi}_s^m \sim \mathcal{U}(0, 2\pi)$ independent of $\mathbf{a}_s^m$ and $f_{0,s}^m$. Hence, here it does not contribute to the solution of (5), although further research may find use for a more informative characterisation of $\boldsymbol{\phi}_s$.

Given a mixed signal of the form in (1), there are a wide range of methods of estimating the values $\mathbf{a}_s$ [27]. In this paper estimates of these values both for training and tracking purposes are obtained with the use of the methods described in Chapter 13 of [18]. Specifically, for a given value of $f_{0,s}^m$, a harmonic set of amplitude estimates obtained by (13.6) of [18] for frame m will be denoted $\hat{\boldsymbol{\rho}}\left(f_{0,s}^m\right)$. The final amplitude and phase estimates at $f_{0,s}^m$ are,

$$\hat{\mathbf{r}}_s^m = \left|(1-\beta)\hat{\boldsymbol{\rho}}\left(f_{0,s}^m\right) + \beta \boldsymbol{\nu}_s^{m-1} \cdot \hat{\boldsymbol{\rho}}\left(f_{back,s}^{m-1}\right)\right|,$$
$$\hat{\boldsymbol{\phi}}_s^m = \text{angle}\left((1-\beta)\hat{\boldsymbol{\rho}}\left(f_{0,s}^m\right) + \beta \boldsymbol{\nu}_s^{m-1} \cdot \hat{\boldsymbol{\rho}}\left(f_{back,s}^{m-1}\right)\right)$$

Here the term $\boldsymbol{\nu}_s^{m-1} = \text{diag}\left(\left[e^{l2\pi i f_{back,s}^{m-1} M}\right]_{l=1}^{L_s}\right)$ is included to allow for the expected phase shift between signal frames. $f_{back,s}^{m-1}$ represents the most likely previous frequency given the current frequency $f_{0,s}^m$, i.e., that which the backpointer points to in the standard Viterbi algorithm. The averaging term is set to $\beta = 0.5$ in this paper. From $\hat{\mathbf{r}}_s^m, \hat{\boldsymbol{\phi}}_s^m$, both log-magnitude estimates $\hat{\mathbf{a}}_s^m$ and the deterministic spectrum in (2) may be obtained.

4. Experimental evaluation

4.1. Implementation

The hidden variables of the FHMM described in Section 3, i.e., $f_{0,1}, f_{0,2}$, were discretised into 251 states linearly spaced between $f_{min} = 50$Hz and $f_{max} = 300$Hz, i.e., at intervals of 1Hz. $P(f_{0,s}^0)$ was trained via the ground truth data of the training utterances by calculating the fraction of observations made in the data at each discrete pmf value. Similarly, $P(f_{0,s}^m | f_{0,s}^{m-1})$ was trained by calculating the fraction of observations of $f_{0,s}^m$ out of all observations of $f_{0,s}^{m-1}$.

In order to solve (3) and hence (5), $\boldsymbol{\lambda}^m$ must be estimated. A simple constant, $\lambda^m[k] = 1, \forall k, m$ was found empirically to provide good results. This value is also approximately equal to the average spectral magnitude across all k in voiced signal segments. Note that in the application of speech enhancement $\lambda_s[k]$ may also be continuously estimated for each source, via the methods of [28, 12] for example.

The multipitch tracking algorithm described in Sections 2 and 3 makes no allowance for voicing decisions, i.e., it assumes deterministic components are always present. For the experiments here, a voicing decision was made via thresholding of:

$$\nu_c = \frac{p\left(\mathbf{Y}^m \left| \hat{\mathbf{H}}^m, f_{est,c}^m, f_{est,d}^m\right.\right)}{\sum_{f_{0,d}^m = f_{min}}^{f_{max}} p\left(\mathbf{Y}^m \left| \hat{\mathbf{H}}^m, f_{est,c}^m, f_{0,d}^m\right.\right)}, \quad (7)$$

where $c \neq d$. This voicing metric is specific to the case $S = 2$.

4.2. Experimental results

The proposed algorithm's performance is compared here to the algorithm in [8]. These algorithms will be referred to as SD-FHMM (proposed), and GMM-FHMM (reference). Performance is measured in the case of speech/noise mixtures where

[4] When estimates are used as arguments of pdfs, this refers to the pdf of the corresponding random variable. For example, $p(\hat{z})$ denotes the pdf of the random variable z, evaluated at the estimate $\hat{z}$.

Table 1: Results for the proposed and reference [8] algorithms for speech/noise mixtures at a range of SNRs

		E_{01}	E_{10}	E_{02}	E_{20}	E_{12}	E_{21}	E_{fine}	E_{perm}	E_{gross}	E_{Tot}
-5dB	SD-FHMM	0.00	0.28	0.00	0.05	5.85	25.77	2.22	0.65	1.33	36.15
	GMM-FHMM	0.00	0.72	0.00	1.18	0.51	22.52	2.29	0.24	0.14	27.62
0dB	SD-FHMM	0.00	1.04	0.00	0.24	4.30	17.29	2.53	0.65	2.42	28.47
	GMM-FHMM	0.00	15.94	0.00	11.00	0.41	39.56	2.89	0.81	0.14	70.74
5dB	SD-FHMM	0.00	2.98	0.00	0.34	3.68	16.06	3.39	0.78	2.47	29.89
	GMM-FHMM	0.00	30.32	0.00	16.19	0.06	41.18	3.98	0.30	0.17	92.18

both speech and noise have deterministic components. Here noise is mixed with speech at SNRs of -5dB, 0dB and 5dB (measured via [29]), of which -5dB is most representative of the signal levels in the training data. The noise used in the experiments here was chainsaw noise recorded from outdoor use of a single chainsaw. Chainsaw noise is known to be composed of both broadband and deterministic components [5] which is somewhat challenging for multipitch tracking algorithms and may be considered explicitly by the SD signal model. The fundamental frequency of a chainsaw covers much of the same range as that of speech. However, its spectral envelope $\mathbf{a}_s$ and the movement of f_0 in time is distinct from speech, and so it provides a good demonstration of how the features in this paper and those in [8] may be used to label a speech or noise source.

So that the results presented are consistent with recent literature in FHMM based speech enhancement, the training data, test data and ground truth pitch data were chosen as close as possible to that in [8]. That is, the speech data consists of the same four speakers from the GRID database [30], using 497 training utterances for each speaker, and the same 3 test utterances for each speaker. All GMMs trained for either algorithm were speaker dependent, i.e., sentences specific to the speakers in each test utterance were used to train each GMM. Ground truth pitch tracks were obtained using the RAPT algorithm [31] for speech data. For noise data an automated pitch tracking algorithm with manual correction was used as suggested in [6]. This was necessary due to the poor performance of the RAPT algorithm on chainsaw noise data.

To measure the performance of each of the algorithms considered, the error metric described in [8] is used. Similar metrics have been used throughout much multipitch tracking literature [6, 22, 8, 32], and so it allows cross-comparison between papers. Specifically, each pitch track obtained is assigned to the source which was used to train the tracking GMMs and pmfs. E_{ij} is the percentage of time frames where i pitch points were misclassified as j pitch points. E_{fine} is the average frequency deviation ($\Delta f^m_{0,s}$) of estimates from reference pitch tracks as a percentage, in frames where $\Delta f^m_{0,s} < 20\%$. E_{perm} is the percentage of time frames where $\Delta f^m_{0,s} > 20\%$ for the assigned pitch track but the estimate was within 20% of the unassigned pitch track. E_{gross} is the percentage of time frames that did not contribute to E_{perm} but for which $\Delta f^m_{0,s} > 20\%$. Finally, E_{tot} is the summation of all other error terms.

The results from the tests conducted are shown in Table 1. It should be noted that in these experiments a gain is applied to the noise signal prior to mixing in order to achieve the desired SNRs. This creates a difference in source levels between the training and test noise data. Whilst both algorithms tested perform well under the -5dB condition, it can be seen that the performance of GMM-FHMM deteriorates completely with more significant changes in the noise source level. Specifically, a significant amount of the noise pitch track is unidentified at these higher SNRs resulting in high rates of E_{10}, E_{20} and E_{21}, whilst the SD-FHMM algorithm shows consistent good performance across all SNRs. The most significant errors of the proposed SD-FHMM algorithm are the voicing errors, E_{12} and E_{21}, which may be improved with a more sophisticated voicing measure than that in (7). Fig. 1 further demonstrates the performance of each of the algorithms. The failure of GMM-FHMM to identify the noise pitch track is shown, and the aforementioned voicing errors of SD-FHMM can be seen.

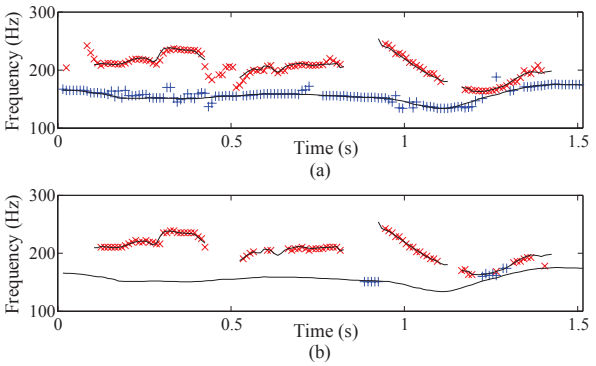

Figure 1: *An example of results obtained from (a) the proposed algorithm and (b) the reference algorithm [8] in a speech/noise mixture of 5dB SNR. In both figures ($\times$) represents a pitch estimate for the speech signal, and ($+$) represents a pitch estimate for the noise signal.*

5. Conclusion

Here an SD-FHMM based multipitch tracking algorithm was proposed for the application of speech enhancement. Specific requirements of such an algorithm were identified to be both reliable labelling of noise/speech sources, and robust performance under conditions where the levels of sources vary. The flexible SD model used allows for a range of features to be exploited for identifying and labelling sources (e.g., $\mathbf{a}_s$ in this paper), and explicit consideration of broadband/stochastic signal components. Integrating this model with an FHMM allows not only the tracking of pitches throughout time, but the ability to exploit this time varying signal information to identify/label sources. The proposed algorithm was compared to a recently developed reference algorithm, also using the FHMM framework. This reference algorithm's use of prior information based solely on the magnitude of a subset of the STFT coefficients is reliable, but heavily dependent on the training data [8]. Furthermore, the reference algorithm ignores a large number of higher frequency spectral coefficients that may be useful in estimating fundamental frequency. The proposed algorithm overcomes these limitations by working directly with complex STFT coefficients and utilising prior information of normalised log-magnitude harmonic envelopes (i.e., $\mathbf{a}_s$). Experiments shown in Section 4 demonstrate that the reference algorithm failed to perform when the level of the noise source varies, in contrast to the proposed algorithm which continued to perform at all SNRs.

6. References

[1] D. Griffin and J. Lim, "Multiband excitation vocoder," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 8, pp. 1223–1235, 1988.

[2] M. Carey, E. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *Spoken Language. Fourth International Conference on*, vol. 3, 1996, pp. 1800–1803.

[3] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *Neural Networks, IEEE Transactions on*, vol. 15, no. 5, pp. 1135–1150, 2004.

[4] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source—filter-based single-channel speech separation using pitch information," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 242–255, 2011.

[5] M. C. McCallum and B. J. Guillemin, "Accounting for deterministic noise components in a MMSE STSA speech enhancement framework," in *Communications and Information Technologies, 12th International Symposium on*, 2012, pp. 174–179.

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

[7] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, and S. Sagayama, "Single and multiple contour estimation through parametric spectrogram modeling of speech in noisy environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1135–1145, 2007.

[8] M. Wohlmayr, M. Stark, and F. 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.

[9] J. Wu, E. Vincent, S. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Multipitch estimation by joint modeling of harmonic and transient sounds," in *Acoustics, Speech and Signal Processing, International Conference on*, 2011, pp. 25–28.

[10] J. Hardwick, C. Yoo, and J. Lim, "Speech enhancement using the dual excitation speech model," in *Acoustics, Speech, and Signal Processing, International Conference on*, vol. 2, 1993, pp. 367–370.

[11] R. Hendriks, R. Heusdens, and J. Jensen, "An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 406–415, 2007.

[12] M. McCallum and B. Guillemin, "Stochastic-deterministic MMSE STFT speech enhancement with general a priori information," *Acoustics, Speech and Language Processing. IEEE Transactions on*, vol. 21, no. 7, pp. 1445–1457, 2013.

[13] F. Bach and M. Jordan, "Discriminative training of hidden markov models for multiple pitch tracking," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 5, 2005, pp. 489–492.

[14] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[15] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *Acoustics, Speech, and Signal Processing, International Conference on*, vol. 2, 1993, pp. 550–553.

[16] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell System Technical Journal*, vol. 27, pp. 109–157, 1948.

[17] B. Rivet, L. Girin, and C. Jutten, "Log-rayleigh distribution: A simple and efficient statistical representation of log-spectral coefficients," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 796–802, 2007.

[18] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[20] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, 1986.

[21] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 1, pp. 21–29, 2001.

[22] M. Wohlmayr and F. Pernkopf, "Multipitch tracking using a factorial hidden Markov model," *Proceedings of the International conference on spoken language processing (ICSLP), Interspeech*, pp. 147–150, 2008.

[23] A. Nadas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 10, pp. 1495–1503, 1989.

[24] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter, *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag, 1999.

[25] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[26] K. Paliwal and L. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[27] S. M. Kay, *Modern Spectral Estimation: Theory & Application*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[28] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 845–856, 2005.

[29] ITU-T, *Objective measurement of active speech level*, ITU-T Recommendation P.56, 2011.

[30] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2005.

[31] D. Talkin, *Speech Coding and Synthesis*. Elsevier Science B.V., 1995, ch. 14: A Robust Algorithm for Pitch Tracking (RAPT), pp. 495–518.

[32] Z. Jin and D. 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.