Improvements to harmonic model for extracting better speech features in clinical applications

Meysam Asgari*, Izhak Shafran

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, Oregon, United States

Received 14 April 2016; received in revised form 10 July 2017; accepted 31 August 2017
Available online 8 September 2017

Abstract
Acoustic properties of speech samples can provide important cues in the assessment of voice pathology and cognitive function. The goal of this study is to develop novel algorithms for robust and accurate estimation of speech features and employ them to build probabilistic speech models for characterizing and analyzing clinical speech. Toward this goal, we adopt a harmonic model (HM) of speech. We overcome certain drawbacks of this model and introduce an improved version of HM that leads us to accurate and reliable estimation of voiced segments, fundamental frequency, HNR, jitter, and shimmer. We evaluate the performance of our improved HM in the context of voicing detection and pitch estimation with other state-of-the-art techniques on the Keele data set. Through extensive experiments on several noisy conditions, we demonstrate that the proposed improvements provide substantial gains over other popular methods under different noise levels and environments. Next, we investigate the utility of developed measures on the speech-based assessment of cognitive impairments including clinical depression and autism spectrum disorder (ASD). Our preliminary results on two clinical tasks demonstrate the promise of our improved HM features in practical applications.

Keywords: Pitch tracking; Voice activity detection; Modified harmonic model

1. Introduction
Analysis of acoustic signals of the human voice has many purposes. Our voice reveals considerable insight into the structure and function of certain organs involved in speech and language production. For instance, sometimes the first symptom of a neurological disorder such as Parkinson’s disease (PD) is a speech impairment (Duffy, 2000). PD can affect all components of speech production including breathing, laryngeal function, and articulation, as well as their coordination for the production of smooth speech. The resulting dysarthric speech often exhibits monotonous pitch, variable speech rate, and a harsh and breathy voice (Darley et al., 1969). Likewise, researchers have shown the effects of psychological disorders, such as depression, in patients’ voices (Low et al., 2011). Speech pathologists have characterized depressed speech as monotonous, mono-loud, and low in range of pitch.
frequency (Moore et al., 2004). Also, a number of studies have shown that emotional arousal considerably influences phonatory and articulatory aspects of the speech production system (Moses, 1954). In addition to PD and depression, autism spectrum disorder (ASD) is another example of a disorder that affects patients’ voices. Children with autism often exhibit unusual pitch and intonation: for example, monotonous pitch, reduced stress, odd rhythm, large pitch range (Hubbard and Trauner, 2007), and even differences in the harmonic structure of their speech (Bonneh et al., 2010).

The severity of the aforementioned diseases is typically assessed subjectively by an expert practitioner and often requires the patient’s presence at the clinic. This assessment is often costly and time-consuming, and can be burdensome in some situations, such as when a patient must undergo frequent reassessments. For several decades now, symptoms observed in the speech of patients with such diseases have motivated researchers to explore alternative approaches based on speech processing techniques. Researchers have measured these symptoms more objectively with the hope of augmenting or simplifying the assessment. It is often cheaper and easier to automatically elicit, record, and analyze speech than to conduct an in-person clinical assessment. Furthermore, speech-based assessment can be remotely administered and can be used to objectively monitor changes over time. Easier methods of assessment, such as automated screening and telemonitoring, can play a crucial role in the early detection of the aforementioned diseases. The main focus of this paper is developing novel algorithms for robust and accurate estimation of pitch-related features, and employing them to build probabilistic speech models for characterizing and analyzing clinical speech. There are a number of approaches in the time and frequency domains to estimate pitch-related features. Time domain methods often ignore frequency and amplitude variations of speech over the analysis frame. On the other hand, the resolution of the short time Fourier transform does not provide the necessary time-frequency resolution to capture the small amount of perturbation observed in, for example, Parkinson’s disease (PD). As an alternative we adopt the harmonic model of speech, which has recently gained considerable attention. This model takes into account the underlying harmonic nature of voiced speech and decomposes it into a harmonic and a non-harmonic component. We overcome certain drawbacks of this model and introduce an improved version of HM that leads us to accurate and reliable estimation of voiced segments, fundamental frequency, HNR, jitter, and shimmer.

2. Speech analysis using the harmonic model

2.1. The harmonic model

The popular source-channel model of voiced speech considers glottal pulses resulting from the rhythmic opening and closing of vocal folds. These pulses are rich in harmonics, and are subsequently shaped by resonances of the oral cavity and the transfer function of the lip radiation. Given that the glottal pulse sequence is rich in harmonics, the resulting voiced sounds can be modeled with a harmonic model containing only non-zero Fourier components at harmonics of the period of the glottal pulses. The harmonic model is a special case of a sinusoidal model where all the sinusoidal components are assumed to be harmonically related; that is, the frequencies of the sinusoids are multiples of the fundamental frequency. This assumption arises from the harmonic nature of the speech signal and reduces the number of parameters in the general sinusoidal model (Stylianou, 1996). Stylianou (1996) introduced a harmonic plus noise model (HNM) for speech analysis and synthesis. The observed voiced signal in HNM is represented in terms of a harmonic component and a non-periodic component related to noise. Speech decomposition using a HNM is useful for applications in speech synthesis, voice conversion, speech enhancement, and speech coding.

2.2. Model expression

To review the model, we adopt the notations from Stylianou (1996). Let \( \mathbf{y} = [y(t_1), y(t_2), \ldots, y(t_N)]^T \) denote the \( N \) speech samples in a voiced frame, measured at times \( t_1, t_2, \ldots, t_N \). The samples can be represented with a harmonic model with an additive noise \( \mathbf{n} = [n(t_1), n(t_2), \ldots, n(t_N)]^T \) modeled by a \( \mathcal{N}(\mu, \sigma_n^2) \) as follows:

\[
\begin{align*}
    s(t) &= a_0 + \sum_{h=1}^{H} a_h \cos(2\pi f_0 h t) + b_h \sin(2\pi f_0 h t) \\
    y(t) &= s(t) + n(t)
\end{align*}
\]
where \( H \) denotes the number of harmonics and \( 2\pi f_0 \) stands for the fundamental angular frequency. The harmonic signal can be factorized into harmonic components that include coefficients of sinusoidal functions, \( a, \beta \), the angular frequency, \( 2\pi f_0 \), and the model order, \( H \).

\[
s(t) = \begin{bmatrix} 1 & A_c(t) & A_s(t) \end{bmatrix} \begin{bmatrix} a_0 \\ a \\ \beta \end{bmatrix}
\]

\[
A_c(t) = [\cos(2\pi f_0 t) \cdots \cos(2\pi f_0 Ht)]
\]

\[
A_s(t) = [\sin(2\pi f_0 t) \cdots \sin(2\pi f_0 Ht)]
\]

\[
a = [a_1 \cdots a_H]^T
\]

\[
\beta = [b_1 \cdots b_H]^T.
\]

Stacking rows of \([1 \ A_c(t) \ A_s(t)]\) at \( t = 1, \cdots, T \) into a matrix \( A \), Eq. (2) can be compactly represented in matrix notation as:

\[
y = A m + n
\]

where the parameter vector, \( m = [a_0, a, \beta] \) contains all the coefficients of sinusoidal functions and \( y = A m \) corresponds to an expansion of the harmonic part of the voiced frame in terms of windowed sinusoidal components. We also define a parameter vector, \( \Theta = [f_0, m, \sigma_n^2, H] \), that summarizes unknown parameters of the model.

### 2.3. ML estimation of model parameters

Assuming the noise samples, \( n \), are independent and identically distributed random variables with a zero-mean Gaussian distribution, the likelihood function of the observed vector, \( y \), given the model parameters can be formulated as the following equation.

\[
L(\Theta) = \log p(y|\Theta)
\]

\[
= -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2} \| y - \Lambda m \|^2
\]

where \( N \) denotes the number of samples in the observed vector, \( y \). Assuming that unknown parameters are independent with respect to each other, the parameters of vector \( m \) (while the other parameters are kept constant) can then be estimated by a maximum likelihood (ML) approach.

\[
\hat{m}_{ML} = (\Lambda^T \Lambda)^{-1} \Lambda^T y
\]

Substituting \( \hat{m} \) into Eq. (4), the likelihood function can be written as follows:

\[
L(f_0, \hat{m}, \sigma_n^2) = -\frac{N}{2} \log(2\pi\sigma_n^2)
\]

\[
-\frac{1}{2\sigma_n^2} \| y - \Lambda(\Lambda^T \Lambda)^{-1} \Lambda^T y \|^2
\]

\[
= -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2} (y - \Gamma)^T (y - \Gamma)
\]

where \( \Gamma = \Lambda(\Lambda^T \Lambda)^{-1} \Lambda^T \) is the projection matrix into the subspace spanned by the columns of \( \Lambda \). The noise variance is maximized by taking the derivative of Eq. (6) with respect to \( \sigma_n^2 \) and making it zero.

\[
\frac{\partial}{\partial \sigma_n^2} L(f_0, \hat{m}, \sigma_n^2) = -\frac{N}{2\sigma_n^2} + \frac{1}{2\sigma_n^2} y^T (I \Gamma) y
\]

\[
\hat{\sigma}_n^2 = -\frac{1}{N} y^T (I \Gamma) y
\]
where $\mathbf{I}$ denotes the identity matrix. This can be simplified further by substituting the estimate of $\hat{\sigma}^2_n$ into the likelihood function:

$$
\mathbf{L}(f_0, \hat{\mathbf{m}}, \hat{\sigma}^2_n) = -\frac{N}{2} \log\left(\mathbf{y}^T (\mathbf{I} - \Gamma) \mathbf{y}\right) - \frac{N}{2} \left(1 + \log\left(\frac{2\pi}{N}\right)\right)
$$

$$
= -\frac{N}{2} \log(\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \Gamma \mathbf{y}) + C
$$

where $C$ is a constant factor does not affect the likelihood. It can be shown that $\Gamma$ is a symmetric matrix where $\Gamma^T = \Gamma$. Also, under the harmonic model, the reconstructed signal $\hat{s}$ is given by $\hat{s} = \mathbf{A} \mathbf{m}$ and the noise component is obtained by subtracting the reconstructed signal from the original speech signal, $\hat{n} = \mathbf{y} - \hat{s}$. Now, using the properties of $\Gamma$, we can rewrite the reconstructed signal and the likelihood function in Eq. (8) as follows:

$$
\hat{s} = \Gamma \mathbf{y}
$$

$$
\mathbf{L}(f_0) = -\frac{N}{2} \log\left(\mathbf{y}^T \mathbf{y} - (\Gamma \mathbf{y})^T (\Gamma \mathbf{y})\right) + C
$$

$$
= -\frac{N}{2} \log(\| \mathbf{y} \|^2 - \| \hat{s} \|^2) + C
$$

$$
= -\frac{N}{2} \log(\| \hat{n} \|^2) + C.
$$

Maximizing the likelihood function in the above equation is equivalent to minimizing the energy of the residual noise component. The energy of the residual part is minimum when the reconstructed signal is fitted well to the original signal. Thus, pitch frequency can be estimated by seeking the frequency that maximizes the energy of the reconstructed signal over a set of discrete $f_0$ values ranging from $f_{0\min}$ to $f_{0\max}$.

$$
\hat{f}_{0\text{ ML}} = \arg \max_{f_0} \hat{s}^T \hat{s}
$$

2.4. MAP estimation of model parameters

The ML estimate of model parameters can potentially over-fit the data. We propose to improve the robustness of the model by exploiting the fact that model parameters cannot vary arbitrarily across frames from the same speaker. Additionally, the physical shape of vocal tract limits the variation of harmonic coefficients into a bounded subspace, and they are not allowed to vary in an arbitrary subspace. From a Bayesian point of view, it is equivalent to imposing a constraint over the space of harmonic components in the model. The maximum a posteriori estimate of the model parameters can be factorized as follows:

$$
\hat{\Theta}_{\text{MAP}} = \arg \max_{\Theta} p(\Theta|\mathbf{y})
$$

$$
= \arg \max_{\Theta} p(\mathbf{y}|\Theta)p(\Theta).
$$

The likelihood of a voiced frame, $p(\mathbf{y}|\Theta)$, is estimated from Eq. (4). For the simplicity, the prior term, $p(\Theta)$, is factorized as $p(\Theta) = p(\mathbf{m})p(\omega)$ where $p(\omega)$ is a uniform distribution from 50 to 500 Hz. In our data, we observe that the coefficients of the harmonic estimated independently per frame using Eq. (5) are approximately Gaussian, as illustrated in Fig. 1.

Hence, we model the prior $p(\mathbf{m})$ as a multivariate Gaussian distribution $\mathcal{N}(\mu_m, \Sigma_m)$. Since the likelihood and the prior are in the same distribution family, Gaussian, the prior is a conjugate prior and we can obtain a closed form MAP estimate by differentiating Eq. (11).

$$
\frac{\partial}{\partial \mathbf{m}} \log p(\Theta|\mathbf{y}) = \frac{2 \Sigma^{-1}_m}{\sigma^2_n} (\mathbf{y} - \mathbf{A} \mathbf{m}) + 2 \Sigma^{-1}_m (\mu_m - \mathbf{m}).
$$

The derivative is set to zero and the closed form analytical expression for the MAP estimate can be computed.

$$
\hat{\mathbf{m}}_{\text{MAP}} = (\mathbf{A}^T \mathbf{A} + \sigma^2_n \Sigma^{-1}_m)^{-1}(\mathbf{A}^T \mathbf{y} + \sigma^2_n \Sigma^{-1}_m \mu_m).
$$
In practice, for each utterance we compute the per-frame estimate of the model parameters, and from those estimates we estimate the prior distribution $N(\mu_m, \Sigma_m)$. Note that the Bayesian estimate of all the model parameters is significantly more complex and requires expensive numerical approximations compared to our simpler MAP estimate with its closed form analytical solution (Godsill and Davy, 2002). The MAP estimate derived in a related previous work smooths the likelihood using a first order HMM transition model and hence differs from our approach (Tabrikian et al., 2004). The first order HMM is learned from the training data with manual pitch annotations and hence may not generalize to test utterances in the clinical domain where the range and transitions may differ considerably across pathology.

2.5. Model order selection

Another problem with the harmonic model is the need to specify the number of harmonics considered. This is typically not known a priori and the optimal value can be different in different noise conditions. Godsill and Davy (2002) proposed a sampling-based method for estimating the number of harmonics. Their approach is based on Monte Carlo sampling and requires computationally expensive numerical approximations. Mahadevan employs the Akaike information criterion (AIC) for tackling the problem of model order selection (Mahadevan and Espy-Wilson, 2011). The AIC attempts to make a balance between the goodness of fit of the model and the model complexity by adding a penalty term to the likelihood. Here, we follow a Bayesian approach, trying to maximize the likelihood function given by:

$$\hat{H} = \arg \max_H p(y, \Theta_H)$$

(13)

where $\Theta_H$ denotes the model constructed by $H$ harmonics. The likelihood increases as a function of increasing model order and often leads to over-fitting. We adopt the Bayesian information criterion (BIC) as a model order selection criterion, where the increase in the likelihood is penalized by a term that depends on the model complexity or the number of model parameters. The BIC has been widely used for model order selection in machine learning problems, such as linear regression and time series (Kadane and Lazar, 2004; Hurvich and Tsai, 1989). The formal definition of BIC is shown in Eq. (14)

$$BIC(\mathcal{M}) \approx -2L(\theta) + \mathcal{M} \log N$$

(14)

where $L$ denotes the likelihood function and $\theta$ denotes parameters of the model. Also, $\mathcal{M}$ and $\mathcal{N}$ are the number of free model parameters and observations, respectively (Wit et al., 2012). Specifically for the harmonic model, Christensen et al. (2011) proposed a MAP-based solution to the above BIC metric shown in Eq. (15). The derivations of the proposed method are somewhat lengthy and we refer readers to their book (Christensen and Jakobsson, 2009).

$$BIC(H) \approx -2L(\Theta_H) + H \log N + \frac{3}{2} \log N.$$  

(15)
In the above equation, the optimal model depends on the number of data points $N$, as well as the number of harmonics assigned to the analysis window. For each frame of speech, the optimal model order is selected by seeking the number of harmonics (ranging from $H = 2, 3, \ldots, H_{\text{max}}$) that minimizes the BIC metric in Eq. (15). Frame-level optimization of model order can be sensitive to noise conditions and lead to over-fitting and under-fitting, leading to the problem of pitch halving and doubling. Instead, to improve robustness against under-fitting and over-fitting, we compute the model order by averaging across frames. We address the problem of halving and doubling further in Sections 2.11 and 2.8. Fig. 2 shows the comparison between the average frame-level score computed by ML and BIC metrics as a function of number of harmonics. In this case, seven harmonics appear to be an optimal trade-off between increasing the likelihood and maintaining a parsimonious model.

### 2.6. Voicing detection

The harmonic structure of voiced speech is preserved even in adverse noisy conditions, and this is a powerful feature for segregating it from the unvoiced and noise signals. Detecting voiced frames can be cast into a hypothesis test problem, in which true and null hypotheses, $H_1$ and $H_0$, are defined as follows:

$$L(H_1) = -\frac{N}{2} \log (\| y \|^2 - \| \tilde{s} \|^2) + C$$

$$L(H_0) = -\frac{N}{2} \log (\| y \|^2) + C$$

where the constant factor $C$ effects both likelihoods equally and is dropped. By comparing the likelihoods, we classify a frame as either voiced or unvoiced. Now that we can compute the probability of observing a voiced or unvoiced frame, the frame-level scores are smoothed to obtain a segment-level decision using a hidden Markov model (HMM), as in Tabrikian et al. (2004). Specifically, this is achieved using an HMM with two states, the voiced ($v$) and the unvoiced ($u$) states, whose observation probabilities are modeled using Eq. (16).

### 2.7. Segmental pitch tracking using harmonic model

As we showed earlier in Eq. (10), pitch frequency can be estimated by maximizing the energy of the reconstructed signal. On the other hand, the rate of pitch variation is inherently limited by the motion of the articulators in the mouth during speech production (Xu and Sun, 2002) and hence does not vary arbitrarily between adjacent frames. This smoothing constraint can be enforced using a first order Markov dependency between pitch estimates of successive frames. Adopting the popular hidden Markov model framework, the estimation of pitch over utterances can be formulated as follows. Let $Y = \{y_0, \ldots, y_M\}$, and $F_0 = \{f_0^{(0)}, \ldots, f_0^{(M)}\}$ be $M$ length sequences of observed frames.
and candidate pitch estimates, respectively. The observation probabilities are assumed to be independent given the
hidden states or candidate pitch frequencies here and are expressed as follows:

\[
p(Y|F_0) = \prod_{i=1}^{M} p(y_i|f_0^i).
\]  

Then, following Bayesian’ rule, the estimation of \( F_0 \) is obtained as follows:

\[
\hat{F}_0 = \arg \max_{F_0} \{ p(Y|F_0)p(F_0) \}.
\]  

A zero-mean Gaussian distribution with a known variance over the pitch difference between frames is a reasonable
approximation for the first order Markov transition probabilities (Tabrikian et al., 2004).

\[
p(F_0) = p(f_0^{(1)}, f_0^{(2)}, \ldots, f_0^{(M)})
\]

\[
= p(f_0^{(1)}) \prod_{i=2}^{M} p(f_0^{(i)}|f_0^{(i-1)})
\]

\[
p(f_0^{(i)}|f_0^{(i-1)}) \sim \mathcal{N}(f_0^{(i)}; f_0^{(i-1)}, \sigma_t)
\]

where \( p(f_0^{(1)}) \) is the prior probability of \( f_0 \) at the first frame, and \( \sigma_t^2 \) is the variation of the pitch transition function.

Putting all this together and substituting the likelihood from Eq. (10), the pitch over an utterance can be estimated
as follows:

\[
\hat{F}_0 = \arg \max_{F_0} \left\{ \sum_{i=0}^{M} \log \mathcal{N}(f_0^{(i)}; f_0^{(i-1)}, \sigma_t^2) \right\}.
\]  

Thus, the estimation of pitch over an utterance can be cast as an HMM decoding problem where its states represent
the possible discrete values of \( f_0 \), and can be efficiently solved using the Viterbi algorithm.

2.8. Pitch halving and doubling

The most common errors in pitch detection algorithms are pitch halving and doubling, mostly known as gross
pitch error (GPE). They often occur due to the strong subharmonics located in the range of pitch frequency; in time
domain, their counterpart are the alternating cycles that appear in both amplitude and period of the speech signal.
These alternating cycles can be found in either disordered or noisy voices. Like in other pitch detection algorithms,
harmonic models suffer from pitch halving and doubling too. The harmonics of \( f_0/2 \) (halving) include all the harmonics
of \( f_0 \). Similarly, the harmonics of \( 2f_0 \) (doubling) are also the harmonics of \( f_0 \). The true pitch, \( f_0 \), may be confused
with \( f_0/2 \) and \( 2f_0 \), depending on the number of harmonics considered and the noise.

In many conventional algorithms, the errors due to halving and doubling are minimized by heuristics such as limiting
the range of allowable \( f_0 \) over a segment or an utterance. This requires prior knowledge about the gender and age of the speakers. Alternatives include median filtering and constraints in Viterbi search (Talkin, 1995), which remain unsatisfactory and unsuitable for clinical applications.

We propose a method to capture the probability mass in the neighborhood of the candidate pitch frequency. The
likelihood of the observed frames falls more rapidly near candidates at halving \( f_0/2 \) and doubling \( 2f_0 \) than at the true pitch frequency, \( f_0 \). This probability mass in the neighborhood can be captured by convolving the likelihood function with an appropriate window.

Fig. 3 illustrates the problem of pitch halving and demonstrates our solution for it. The top plot represents a
voiced frame of clean speech contaminated with the babble noise at 10 dB SNR, and in the middle plot, the dotted
line shows its estimated likelihood function. A maximum of this function will erroneously pick the candidate \( f_0/2 \) as
the most likely pitch candidate for this frame. However, notice that the function has a broader peak at \( f_0 \) than at \( f_0/2 \). Moreover, the area under the curve of the likelihood function, i.e. the energy of the reconstructed signal, is higher in the neighborhood (represented by \( G \) in following equation) near \( f_0 \) than \( f_0/2 \) as shown in middle plot of this figure.

\[
\sum_{f \in \mathcal{G}(f_0)} E(f) \gg \sum_{f \in \mathcal{G}(f_0/2)} E(f).
\]
Convolving the reconstructed signal with an appropriate window will result in suppressing the peak at $f_0/2$ and raising the peak at $f_0$ due to the higher energy mass in its neighborhood. The solid line in the bottom plot of this figure shows the result of convolving the likelihood function with a hamming window. In our experiments, we employed a hamming window with a length of $f_{0\min}/2$ where $f_{0\min}$ is the minimum pitch frequency set to be 50 Hz in our experiments.

Fig. 3. The top plot is a voiced frame of speech contaminated with the babble noise at 10 dB SNR. The solid and dash lines in the bottom plot illustrate original and smoothed likelihood functions, respectively. The likelihood function has maxima near $f_0$ and $f_0/2$. The middle plot shows the higher probability mass near $f_0$ shown by the red patch, than $f_0/2$ shown by blue patch. The bottom plot shows how locally smoothing the likelihood solves the pitch halving problem. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
experiments. The locally smoothed likelihood shows a relatively high peak at the true pitch frequency \( f_0 \) compared to \( f_0/2 \), thus overcoming the problem of pitch halving.

### 2.9. Empirical evaluations

We evaluate the performance of our proposed method on a task of estimating pitch frequency on the Keele dataset that contains 10 phonetically balanced audio files from 10 speakers, 5 male and 5 female. This dataset provides a reference pitch and voicing labels obtained from the simultaneously recorded laryngograph signal. For evaluation, we exclude frames for which the voicing label in the corpus is uncertain. The speech was recorded in noise free conditions. To test the robustness of our algorithm, we contaminated the files at different noise levels ranging from 0 dB to 20 dB in several noisy environments including restaurant, subway, white, car, street, exhibition, babble, and airport taken from the Aurora noise dataset (Hirsch and Pearce, 2000). For adding the noise signal, we used Filtering and Noise adding Tool (FaNT) (Hirsch and Pearce, 2000) with the telephone speech characteristics configuration using the G.712 filter, a narrow-band telephone speech bandpass filter with a flat frequency response between approximately 300—3400 Hz. This filtering makes the task of pitch estimation more challenging than in the full-band scenario due to the spectral attenuation at harmonics below 300 Hz.

We assessed performance of the pitch trackers using the following measures (Drugman and Alwan, 2011): gross pitch error (GPE), which is defined as the percentage of \( f_0 \) estimates that deviate more than 20% from the ground truth; and fine pitch error (FPE), which is the mean absolute error computed for estimates that are below than 20% of reference \( f_0 \).

\[
GPE = \frac{Z}{M} \times 100
\]

\[
FPE = \frac{1}{M} \sum_{i=1}^{M} |\hat{f}_0^{(i)} - f_0^{(i)}_{gt}| \quad (22)
\]

where \( Z \) denotes the number of frames with more than 20% of deviation from the ground truth in their pitch estimates. Also, \( \hat{f}_0^{(i)} \) and \( f_0_{gt}^{(i)} \) denote the estimated and ground truth pitch frequencies in the \( i^{th} \) frame. We compared the performance of our proposed method with the following pitch estimation methods – (a) STRAIGHT-TANDEM (S-T), based on the fixed-point analysis on a modified power-spectrum (Kawahara et al., 2008); (b) YIN, a template matching method with the autocorrelation function in time-frequency domain and ad hoc post-processing (De Cheveigné and Kawahara, 2002); and (c) SHR, a method based on Subharmonics to Harmonic Ratio (Sun, 2002). In all cases, the search for optimal pitch frequency was performed over a range from 50 Hz to 500 Hz and the frame rate was fixed to 100 frames per second. Note that the performance of pitch trackers is compared using only the frames corresponding to the reference voiced frames.

### 2.10. Experimental results

In noisy conditions, Table 1 reports the average errors over all SNR bins ranging from 0 dB to 20 dB in terms of gross pitch error (GPE). On both clean and noisy speech, the harmonic model with smoothed likelihood (HM-SL) clearly outperforms all other approaches in terms of GPE except in the white noisy condition, where the smoothing appears unnecessary and the HM outperforms others.

Table 2 reports average errors in terms of fine pitch error (FPE), in which the HM approach outperforms others except for SHR in the restaurant noisy condition. As it is clear in Table 2, HM-SL has a comparable performance to the HM. This may be explained by the fact that smoothing of the likelihood score may reduce the precision of the harmonics.

Fig. 4 summarizes the overall gross pitch error and fine pitch error across all the noise conditions. The proposed model (HM-SL) substantially outperforms the other popular pitch trackers in gross pitch error under all levels of noise conditions. The performance of HM-SL is also better than HM, which shows that the smoothing contributes to the performance gains. The model also performs better in fine pitch error when the noise level is high. At low noise
levels, the proposed model degrades the fine pitch estimate, which is not entirely surprising and is due to the smoothing. In fact, at high SNRs the standard HM is sufficient and smoothing is not necessary.

2.11. Model sensitivity

In order to measure the sensitivity of the proposed pitch tracker to variation of the model order, we evaluated the overall GPE under both clean and noisy conditions as a function of number of harmonics, $H$. For the noisy condition, we averaged the GPE over all noise types. Fig. 5 indicates that under clean conditions, the model is robust to model order variations. However, as SNR increases, GPE varies as a function of model order. We computed average frame-level BIC for several numbers of harmonics, ranging from $H = 2, \ldots, H = 9$, on held-out data. In this experiment, ten percent of recordings (noisy and clean) were randomly selected and seven harmonics appeared to be an optimal value across clean and noisy conditions at different SNRs.

The same experiments were performed to evaluate the model sensitivity to variance of the transition function in adjacent frames. The results, illustrated in Fig. 5, show that the optimal variance in transition function, about 2 Hz, is fortunately not sensitive to a range of SNRs.

3. Clinical speech applications

In this section, we utilize our improved harmonic model and extract pitch related features, along with standard acoustic measures, and investigate their effectiveness in clinical applications. We employ supervised machine learning algorithms, such as support vector regression (SVR), to learn probabilistic speech models for automatically characterizing speech impairments in Parkinsons disease, autism spectrum disorder (ASD), and clinical depression. Note
that although our features are just as easily employed in conjunction with deep neural networks, the limited clinical data available to us precludes us from investigating this extension.

3.1. Inferring clinical depression from speech and spoken utterances

Clinical depression is a common mental disorder that negatively affects person’s health, mood, thoughts, behavior, work, family, and ability to function in everyday life (Pine et al., 1999; Caspi et al., 2003).

Our corpus for this study was collected by Oregon Research Institute (ORI) and consists of video recordings of adolescent subjects interacting with their families. Subjects were asked to participate in three different 20-minute interactions with their parents: event-planning interaction (EPI), problem-solving interaction (PSI), and family consensus interaction (FCI) (Hops, 1995). All interaction were administrated by a trained interviewer in a quiet room at ORI. The recordings were collected from 148 subjects, including 98 females and 50 males from 14 to 18 years old. Of these subjects, based on clinical assessment, 71 adolescents (50 females and 21 males) were diagnosed as depressed and 77 individuals (48 females and 29 males) were healthy controls.

As a clinical reference, the severity of subjects’ conditions were coded using the living-in-family-environment (LIFE) coding system (Hops, 1995). The LIFE coding system was developed for the assessment of behavioral characteristics of individuals with depressive disorders. In the LIFE coding system, behavior is coded based on verbal, nonverbal, and paraverbal behavior, so codes do not necessarily imply that the subject is speaking. A group of psychologists at ORI coded subjects’ verbal content and emotional state using 27 content codes and 10 affect codes available on the LIFE coding system. Subjects’ behavior were marked by six categories: angry, dysphoric, happy,
neutral, end, and other, for each second in family interaction sessions. For the purpose of our experiments, we extracted audio from the video recordings and converted them from stereo to mono channel format. Then, speech segments annotated with angry, dysphoric, and happy tags were chopped and concatenated for manual transcription. There was noise in the annotations as they were not always completely aligned with the utterances. We asked annotators to manually identify the speaker at each segment to alleviate noise in the labels.

3.1.1. Textual features

In order to gauge the effect of speech content in clinical depression, we extracted textual features from manual transcripts. To extract features from text, we used a published table of valence and arousal ratings by Warriner et al. (2013) to tag each word in an utterance with an arousal and a valence rating and compute their per-utterance mean, standard deviation, minimum, and maximum. For missing words we imputed valence and arousal by randomly drawing 5 words from the table and computing their average.

3.1.2. Speech features from openSMILE

In our experiment, we used a broad range of features to capture speech clues associated with clinical depression. For our baseline system, we adopted the baseline features defined in Interspeech 2010 Paralinguistic Challenge (Schuller et al., 2010) using openSMILE toolkit (Eyben et al., 2010).

The features, comprised of 1596 components, can be broadly categorized into three groups: (1) loudness related features such as RMS energy and PCM loudness, (2) voicing related features like pitch frequency, jitter, and shimmer, and (3) articulatory related features such as mel-frequency cepstral coefficients and line spectral frequencies.
The above configuration provides 38 features along with their derivatives to form the frame-level acoustic features. The derivatives allow us to capture local dynamics of pitch and other features. Using 21 standard statistical functions including min, max, mean, skewness, quartiles and percentile, the features computed at the frame-level were summarized into a global feature vector of fixed dimension $21 \times 38 \times 2 = 1596$ for each recording.

### 3.1.3. Speech features from harmonic model

Alternatively, we extracted prosodic features using the harmonic model of speech described earlier. Briefly, we extract 25 ms long frames using a Hamming window with a 10 milliseconds shift. We first detect voiced frames robustly by calculating the likelihood of voicing under the harmonic model. The voicing decision at the segment level is computed by formulating a one-state hidden Markov model (HMM) (Asgari et al., 2012). The state could either be voiced or unvoiced, with likelihood given by the per-frame harmonic model. The transition model consists of a simple zero-mean Gaussian. We compute voicing decisions over the utterance using Viterbi alignment. Subsequently, we compute various voicing related features for voiced frames, including pitch frequency, HNR, H12, jitter, shimmer, and harmonic coefficients. For further detail, we refer readers to Asgari and Shafran (2010). These pitch-related features are combined with standard features including energy, spectral entropy, and MFCCs. For unvoiced frames, we just compute energy, spectral entropy, and MFCC features. Features extracted from voiced regions tend to differ in nature compared to those from unvoiced regions. These differences were preserved and features were summarized in voiced and unvoiced regions separately. Per-utterance features are computed by applying standard summary statistics such as mean, median, variance, minimum and maximum to the per-frame voiced (unvoiced) features, generating a 192-dimensional per-utterance feature vector.

### 3.1.4. Experiments

We compared the performance of a SVM classifier on our data using 10-fold cross-validation, setting all model parameter using 9 of the 10 sets (133 subjects) as training set, and using the tenth ones (15 subjects) only for reporting the performance on classifying depressed versus control subjects. Table 3 reports the performance of classifiers trained on different feature sets. The SVM classifier with several kernel functions, including linear, polynomial, and radial basis functions, was employed from the open-source Scikit-learn toolkit (Pedregosa et al., 2011). Parameters of the optimal classifier were determined via grid search and cross validation on the training set. The best performance among all models was obtained with the linear kernel except for the openSMILE feature set with RBF kernel. As it is shown in Table 3, speech features, extracted from both openSMILE and harmonic model, perform significantly better than chance with a p-value of less than 0.01, according to cross-validated paired t-test (Dietterich, 1998) and denoted by $(\dagger)$ in the table.

Also, Table 4 indicates that incorporating textual features by concatenating them with speech features results in an additional improvement. Spoken words contain information for detecting depression as expected and use of solely textual features performed significantly better than chance.

### 3.1.5. Effectiveness of family interaction sessions

In this section, we separately examine the influence of each family interaction session in classifying the depressed speech. As we mentioned earlier, there are three types of family interactions: (1) EPI, (2) PSI, and (3) FCI. These tasks evoke different emotional states in adolescents during the family interaction and potentially reveal different aspects of their behavior. For instance, PSI tends to elicit the conflictual behavior of adolescents when interacting with their parents. Table 5 reports the performance of SVM classifiers. The results indicate that features extracted from PSI are most effective in this task. The performance of the classifier trained on features extracted from the harmonic model is significantly better than chance with p-value of less than 0.01. Our analysis of the results

<table>
<thead>
<tr>
<th>Features</th>
<th>Classification accuracy</th>
<th>SVM Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>52.4</td>
<td>–</td>
</tr>
<tr>
<td>openSMILE</td>
<td>66.7(\dagger)</td>
<td>RBF</td>
</tr>
<tr>
<td>Harmonic Model</td>
<td>69.8(\dagger)</td>
<td>Linear</td>
</tr>
</tbody>
</table>
show that the FCI is not an effective task for evoking clues associated with clinical depression of adolescents, and that features extracted from this family session are poor predictors of its severity (Asgari et al., 2014).

### 3.2. Detecting and diagnosing autism spectrum disorders

Finally, we show the advantage of our proposed pitch tracker on the *Autism Sub-Challenge* of Interspeech 2013 (Schuller et al., 2010). The challenge consists of two tasks: (1) a binary ‘Typicality’ classification task with classes – TYPically developing (TYP) and ATYPically developing (ATY), and a four-way ‘Diagnosis’ task for classifying children into 4 categories – TYP, pervasive developmental disorders (PDD), pervasive developmental disorder not otherwise specified (PDD-NOS), and specific language impairment such as DYSphasia (DYS). Empirical evaluation were performed on the “Child Pathological Speech Database” (CPSD) (Ringeval et al., 2011) collected from 99 children, ages 9 to 18, through two hospitals located in Paris, France. This dataset provides 2542 short speech utterances collected for assessing children’s abilities in imitation of different types of prosody contours. Based on the prosodic dependencies of French language, sentences carry out 4 intonation types including *descending, falling, floating, and rising*. Subjects were asked to read 26 phonetically easy sentences which were recorded in separate files. As a clinical reference, the severity of subjects was measured by clinicians using the DSM-IV criteria (Guze, 1995), where 35 of these children showed PDD: either Autism Spectrum Condition (ASC, 12 children), specific language impairment (SLI, 13 children) or PDD-not otherwise specified (PDD-NOS, 10 children). The corpus includes rich annotation such as speaker meta-data, orthographic transcript, phonemic transcript, and segmentation. Also, the corpus treats sentences read by the same speaker as independent samples partitioned randomly into test, development, and training sets.

For the experiments, we estimated the parameters of a harmonic model for detecting the voiced frames and estimating the pitch frequencies ($f_0$). Then, from the fundamental frequencies and the reconstructed noise-free signal, we computed other voicing related features such as Harmonic-to-Noise Ratio (HNR), shimmer, and jitter. The features computed at the frame-level needed to be summarized into a global feature vector of fixed dimensions for each read sentence. Each feature was summarized across all frames from the voiced segments in terms of standard distribution statistics such as mean, median, variance, minimum and maximum. We also computed the covariance matrix (upper triangular elements) of frame-level feature vectors over voiced segments to capture interaction between features. The resulting per-sentence voice quality feature vector was later augmented by per-sentence energy, spectral, and cepstral related features provided from the baseline. For the comparison and as a baseline system, we replicated the experiments using speech features provided in the challenge. For more detail regarding the baseline frame-level features and also statistical functions applied to those features, we refer readers to the paper published by the

| Speech Features | Classification Accuracy | | |
|-----------------|-------------------------|---|---|---|
| Chance          | 49.2                    | 49.2 | 49.2 |
| openSMILE       | 60.0                    | 64.7 | 56.0 |
| Harmonic model  | 66.1                    | 71.4 | 57.6 |

<table>
<thead>
<tr>
<th>Features Classification accuracy</th>
<th>SVM Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>52.4</td>
</tr>
<tr>
<td>Text</td>
<td>64.1</td>
</tr>
<tr>
<td>openSMILE+Text</td>
<td>66.8</td>
</tr>
<tr>
<td>Harmonic Model+Text</td>
<td>74.8</td>
</tr>
</tbody>
</table>

**Table 4**

Comparison of performance of SVM classifier using different combination of textual and acoustic features for classifying clinical depression in adolescents.

**Table 5**

Effect of family interaction sessions on classifying clinical depression in adolescents.
organizers of the challenge (Schuller et al., 2013). Inspired from the baseline, Support Vector Machines (SVM) and Support Vector Regression (SVR) with linear kernel were employed for the binary classification and four-way regression tasks, respectively. Table 6 reports UAR evaluated from baseline features derived from openSMILE and proposed features for both diagnosis and typicality tasks. Note that model parameters in the SVM and SVR models are shared between both assessments. As is clear in results, substituting proposed voicing related features (derived by harmonic analysis) with voicing features provided in the baseline feature vector significantly improves UAR in both tasks. Our system gave the best results in the challenge (Asgari et al., 2013).

4. Conclusion

Pitch-related features play an important role in characterizing pathologies such as Parkinson’s disease, clinical depression and autism spectral disorder (ASD). Of the available pitch estimators, only the harmonic model exploits the inherent harmonic structure of the voiced speech and has model parameters which can be inferred on-the-fly without the need to learn a model on training data. This is particularly attractive since creating manually annotated pitch data for different pathologies which include a wide range of representative patients is labor intensive and expensive. Thus we adopt the harmonic model for our work. We address a few drawbacks of the model and demonstrate improvements in accuracy and robustness over a range of SNRs. Subsequently, we employ our model on two clinical tasks and extract pitch-related features. On the task of detecting clinical depression, our features give about 9% relative improvement over those from openSMILE. On the 2013 Interspeech Autism challenge, compared to a state-of-the-art baseline provided by the organizers, our features improve the performance, measured in terms of unweighted average recall (UAR), of detecting ASD by 2.3% and of classifying the disorder into four categories by 2.8%. Our system gives the best performance in the challenge. Taken together, we show that these features are useful in practical applications and we hope this stimulates further research in automated speech-based assessment of clinical tasks.

References


