IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 27, 2018, accepted May 24, 2018, date of publication May 28, 2018, date of current version June 26, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2841802

Hybrid Autoregressive Resonance Estimation and Density Mixture Formant Tracking Model

MIGUEL ARJONA RAMÍREZ, (Senior Member, IEEE)
Department of Electronic Systems Engineering, Escola Politécnica, University of São Paulo, São Paulo 05508-010, Brazil
e-mail: miguel@lps.usp.br

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo under Grant 2015/25512-0.

ABSTRACT A novel formant tracker is proposed using the mixture models of *t* densities (tMMs) for vocal tract resonance frequencies estimated with a hybrid linear prediction (HLP) method. The hybrid integer-cycle pitch-synchronous linear prediction (LP) analysis improves the frequency resolution over voiced segments, leading to closer formant estimates than those provided by other LP methods. In conjunction with HLP, formant trajectories are shown to be more nearly tracked by tMMs than by Gaussian density models. Tests with synthetic voiced and whispered speech as well as with an annotated database confirm better performance than either tMM clustering after formant estimation based on different time-frequency representations or tracking after different LP methods.

INDEX TERMS Formant estimation, formant tracking, vocal tract resonance, autoregressive models, t mixture models, tMM.

I. INTRODUCTION

Formant frequencies are very important in human speech perception and are widely used in phonetics. However, other features are commonly used for automatic speech analysis and recognition because they are estimated with straightforward algorithms.

Part of the difficulty in using formant frequencies lies in their multiple definitions. They were originally defined as peaks in the speech spectral envelope but, more recently, vocal tract resonance (VTR) frequencies [1], [2] have come to be considered a better representation. We will adhere to the latter definition in the estimation phase and, in the tracking phase, the formant frequencies will finally be identified with the means of the density models.

Even though formant frequencies often manifest themselves by spectral envelope peaks, there are notable cases when the VTR frequencies are hidden such as when there is the combined effect of open nasal cavities in addition to the oral tract or when formant tracks touch in velar pinch patterns. That is why VTR frequency trackers have previously been proposed using hidden dynamics trackers [2]–[4].

Formant estimation may be based on a transform representation or may use an intermediate model. Transformed parameters that have been used include the LP cepstra [2], pole frequencies of the LP model of the analytic speech signal [5], [6] and Gammatone filterbank signals [7]. Also, parameters related to formant frequencies, like the spectral subband centroids (SSCs), have been proposed for speaker verification [8] and speaker recognition [9] on the grounds that mel frequency cepstral coefficients (MFCCs) are more sensitive to noise than SSCs.

Linear prediction methods range from the autocorrelation method [5] to the covariance method for the complex-valued analytic signal [6]. Given the importance of voiced frames for accurate formant estimation, a variant autocorrelation method has been proposed using electroglottography signals for the extraction of pitch marks [10].

Formant tracking may be achieved by hidden dynamics as mentioned above and also by dynamic programming (DP) [5]. In addition, formant dynamics may be closely tracked by the component means of *t* density mixture models (tMMs) fit to the speech pyknogram, which is a time-frequency representation of the evolving speech signal [11]. Since the power spreads around the mean with long tails, they have found tMMs to provide a much better fit than Gaussian mixture models (GMMs). However, GMMs have been used to represent the power spectral density (psd) when the number of components is adapted by a Dirichlet process mixture model [4].

In the following, we propose a hybrid LP method to estimate vocal tract resonance frequencies and adapted tMMs to identify formant trajectories. Finally, the proposed algorithm is compared with existing formant estimation and tracking algorithms and also the hypotheses used in its construction are put to the test, including the linear prediction method and the density tracker.

II. HYBRID LINEAR PREDICTION

Vocal tract resonances are more significant for voiced speech, where linear prediction (LP) methods should be used with care due to the spectral sampling caused by periodic voiced excitation. Great effort has been spent trying to unveil the true spectral envelope in voiced frames [10]. In particular, the STRAIGHT psd obtained from TANDEM-STRAIGHT spectrograms [12] is free from harmonic artifacts and its stable spectral envelopes are excellent for speech synthesis [13].

In fact, the STRAIGHT power spectral density (psd) $S\left(e^{j\omega}\right)$ could be used to obtain an estimate of the vocal tract transfer function $H(z)$ by using the Wiener-Khinchin theorem to obtain the autocorrelation coefficients

$$R(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S\left(e^{j\omega}\right) e^{j\omega m} d\omega \tag{1}$$

for $m = 0, 1, \ldots, p$ to be used as coefficients in the set of LP autocorrelation equations. Since the STRAIGHT psd is free from harmonic artifacts [13], this algorithm is used for resonance frequency estimation as a comparison in our tests while pitch and voicing estimates from TANDEM-STRAIGHT are also used in the proposed hybrid LP formant estimator.

Other methods that display low sensitivity to harmonic interferences involve the use of the analytic signal as done in [6], which, despite the improved precision of its f_{R1} estimates, causes greater deviations in f_{R3} and f_{R4} estimates. Greater precision estimates for f_{R1} and f_{R2} are also obtained by a time-domain weighting of the square of the prediction error signal in a method targeted for high-pitched speech when the interest is restricted to these first two formants [14]. But we propose a resonance estimator which can provide accurate estimates beyond the second formant to be useful in applications besides linguistics such as speaker identification and emotion recognition. Fortunately, as will be appreciated through estimation results, our method is able to keep up with the good f_{R1} and f_{R2} estimates using the real-valued speech signal as long as the fundamental frequency f_o estimates are precise and the window segmentation is pitch-synchronous.

When selecting the method for LP analysis, it should be considered that, in previous tests, the autocorrelation method has been found to need very precise pitch marks complemented by a compensation of the spectral sampling as in [10]. On the other hand, the covariance method may be modified to account more precisely for the evolution of the voiced speech signal in the short term.

The modified pitch-synchronous covariance analysis allows a simpler solution to spectral envelopes with reduced spectral sampling artifacts and great VTR frequency accuracy in HLP for voiced frames by means of a rectangular window for a pitch-synchronous selection of an integer number of pitch periods for the summation range of the correlation coefficients

$$\varphi_{ij} = \sum_{n=p}^{p+N_v \cdot p_o - 1} s_p(n-i) s_p(n-j), \tag{2}$$

for $j = 0, 1, \ldots, p$ and $i = j, j+1, \ldots, p$, where $s_p\,(n)$ is the preemphasized speech signal

$$s_p(n) = s(n) - \mu s(n-1), \tag{3}$$

where $s(n)$ is the input speech signal, μ is the preemphasis factor to be specified in Section IV, p_o is the current pitch period length, p is the LP order and the number of pitch periods selected within the current L sample long window is

$$N_v = \left\lfloor \frac{L-p}{p_o} \right\rfloor .$$

The use of the correlation coefficients in (2) with the covariance method, in what we will call the pitch-synchronous covariance analysis, effectively prevents the inherent application of the Bartlett or triangular window to the autocorrelation sequence in short-term LP analysis. This provides better resolution in periodicity and sensitivity to true amplitude and waveshape variations.

The correlation coefficients enter the set of LP covariance equations

$$\boldsymbol{\Phi a} = -\boldsymbol{\psi}, \tag{4}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \varphi_{11} & \varphi_{21} & \cdots & \varphi_{p1} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{p1} & \varphi_{p2} & \cdots & \varphi_{pp} \end{bmatrix}$$

and $\boldsymbol{\psi} = \begin{bmatrix} \varphi_{10} & \varphi_{20} & \cdots & \varphi_{p0} \end{bmatrix}^T$ are the $p \times p$ correlation matrix and the $p \times 1$ correlation vector, respectively, while vector $\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}^T$ contains the prediction coefficients.

The solution may be efficiently obtained in the polynomial vector space [15], [16] for the transfer function

$$\begin{aligned} A(z) &= 1 + \sum_{i=1}^{p} a_i z^{-i} \\ &= \prod_{i=1}^{p} \left(1 - \varpi_i z^{-1}\right), \end{aligned} \tag{5}$$

of the inverse filter, whose zeros are ϖ_i, $i = 1, 2, \ldots, p$. They are also poles of the estimated vocal tract transfer function, given in factored form by

$$H(z) = \frac{1}{\prod_{i=1}^{p} \left(1 - \varpi_i z^{-1}\right)}. \tag{6}$$

The complex poles are sorted out for their arguments, which determine the pole frequencies

$$f_i = \frac{\arg \varpi_i}{2\pi} \tag{7}$$

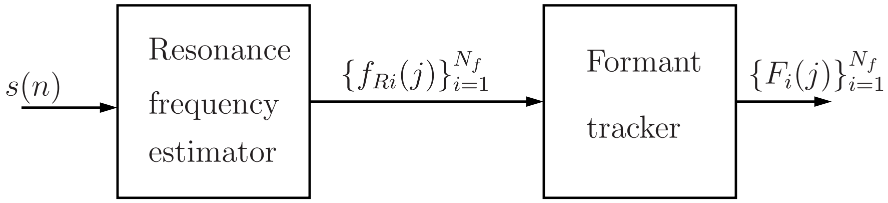


FIGURE 1. Composition of formant algorithm consisting of a resonance frequency estimator and a formant tracker.

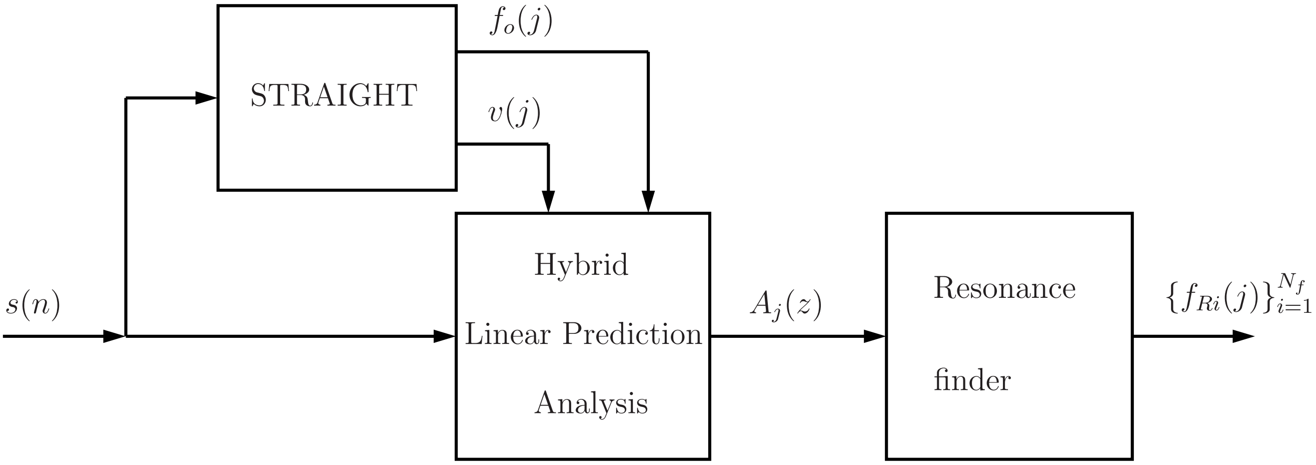

STRAIGHT
$f_o(j)$
$v(j)$
$s(n)$
Hybrid Linear Prediction Analysis
$A_j(z)$
Resonance finder
$\{f_{Ri}(j)\}_{i=1}^{N_f}$

FIGURE 2. Resonance frequency estimator based on hybrid linear prediction analysis.

for the complex poles ϖ_i, $i = 1, 2, \ldots, N_c$, where the number of complex poles is $N_c \leq p$ and they appear in conjugate pairs. Further, from the set of complex pole frequencies, the positive frequencies υ_i are selected. The positive frequencies are denormalized by the sampling frequency F_s and those above a low-frequency threshold $f_{R\,\min}$ are the corresponding VTR frequencies, given by

$$f_{Ri} = \max\left(F_s\upsilon_i, f_{R\,\min}\right), \tag{8}$$

for $i = 1, 2, \ldots, N_f$, where N_f is the number of formants. This work uses at most $N_f = 4$ and sets $f_{R\,\min} = 90$ Hz.

In HLP, unvoiced frames are analyzed with the autocorrelation method of linear prediction for smoother trajectories in this hybrid LP method. The autocorrelation coefficients

$$R(i) = \sum_{n=0}^{L-i-1} s_p(n)s_p(n+i) \tag{9}$$

are computed over each window for $i = 0, 1, \ldots, p$. With the autocorrelation coefficients, the set of LP autocorrelation equations

$$\boldsymbol{Ra} = -\boldsymbol{r} \tag{10}$$

is constructed, where $\boldsymbol{r} = \left[R(1)\ R(2)\ \cdots\ R(p)\right]^T$ and

$$\boldsymbol{R} = \text{toeplitz}\left(\left[R(0)\ R(1)\ \cdots\ R(p-1)\right]^T\right)$$

is a $p \times p$ symmetric Toeplitz matrix.

A comparison of the performance of HLP with those of autocorrelation LP and pitch-synchronous covariance LP is presented in Section IV.

The proposed formant algorithm is to have the structure outlined in Fig. 1, composed of a resonance frequency estimator and a formant frequency tracker, where the intermediate resonance frequencies and the output formant frequency trajectories are produced at the frame rate. By the way, each combination of estimator and tracker will be simply referred to as a formant algorithm. This structure is also valid for most of the formant algorithms to be used for comparison.

The structure of the resonance frequency estimator based on HLP analysis is depicted in Fig. 2, where the linear prediction analysis receives the input speech signal $s(n)$ and also, from STRAIGHT, the frame rate voicing indicator $v(j)$ and the fundamental frequency $f_o(j)$ sequences, delivering the frame rate sequence of analysis filters $A_j(z)$. As a last step, the resonance frequency estimates are found by solving for the zeros of the analysis filter, whose frequencies are further bounded by Eq. (8) above. Next, the evolution of the resonance frequencies must be smoothed for proper formant tracking and the best tracker has been found to be based on t density models as explained in Section III.

III. MIXTURE MODELS OF *t* DENSITIES

As mentioned previously, it has been found that the t probability density function (pdf) fits formant trajectories better

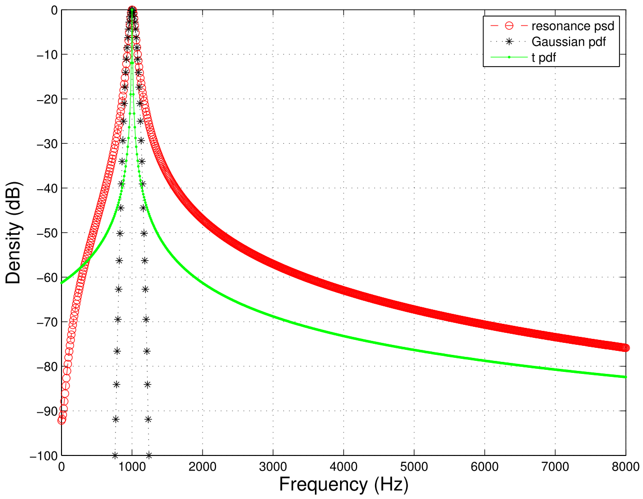


FIGURE 3. Vocal tract resonance psd for resonance frequency $f_R = 1$ kHz and bandwidth $B_R = 100$ Hz, peak normalized Gaussian pdf with variance $B_R^2/4$ and t pdf with degrees of freedom $\nu = 1/4$, both about f_R.

than the Gaussian pdf. In [11], it is speculated that this may be understood by considering the impulse response of a single resonance

$$h(t) = e^{-\alpha_R t}\cos(\Omega_R t)\cdot u(t),$$

with $\Omega_R = 2\pi f_R$ being f_R the resonance frequency and $\alpha_R = \pi B_R$ being B_R the resonance bandwidth. The psd for this resonance is

$$S(j\Omega) = \frac{\Omega^2 + \alpha_R^2}{\left(\Omega^2 - \Omega_R^2 - \alpha_R^2\right)^2 + 4\alpha_R^2\Omega_R^2}. \tag{11}$$

It can be seen in Fig. 3 that this psd has longer tails than the Gaussian pdf and that it may be fit more closely by the t pdf, which also displays polynomial decay. The adjustment is enhanced by parameter ν, the degrees of freedom in its pdf

$$p_t\left(f;\mu,\sigma^2,\nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\cdot\Gamma\left(\frac{\nu}{2}\right)\sigma}\left(1 + \frac{(f-\mu)^2}{\sigma^2\nu}\right)^{-\frac{\nu+1}{2}}, \tag{12}$$

where the mean is $\mu = f_R$ and the variance is $\sigma^2 = B_R^2/4$.

The representation of a tMM density consisting of N_f n-variate density components with parameter set $\boldsymbol{\Psi} = \left\{c_1, c_2, \ldots, c_{N_f}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{N_f}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{N_f}, \nu_1, \ldots, \nu_{N_f}\right\}$ is

$$p_{\,\mathrm{tMM}}(\boldsymbol{f};\boldsymbol{\Psi}) = \sum_{i=1}^{N_f} c_i p_t\left(\boldsymbol{f};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\nu_i\right), \tag{13}$$

where the n-variate t density $p_t(\cdot;\cdot,\cdot,\cdot)$ is given by

$$p_t(\boldsymbol{f};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\sqrt{\pi\nu}\cdot\Gamma\left(\frac{\nu}{2}\right)|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot\left(1 + \frac{(\boldsymbol{f}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+n}{2}} \tag{14}$$

and the nonnegative component weights c_i must add up to unity.

Resonance frequencies, as determined in Section II, are collected for frame j as $\boldsymbol{f}_j = \left[f_{R1,j}\, f_{R2,j}\, \ldots f_{RN_f,j}\right]^T$ and then these column vectors are concatenated over a range in j of N_t estimation windows for the application of the expectation-maximization (EM) algorithm to determine the optimal parameter set $\boldsymbol{\Psi}$ [17, Chapter 7].

Initially, the VTR frequencies are partitioned into N_f clusters in bands about integer kHz values with standard deviations of 100 Hz, equal component weights and unit degrees of freedom [11].

In Section IV, tMMs are favorably compared to GMMs in vowel formant tracking. In the E-step the posterior probability that $\boldsymbol{f}_j$ belongs to the ith component of the mixture using the current parameter values $\boldsymbol{\Psi}$ is computed as

$$\chi_{ij} = \frac{c_i p_t\left(\boldsymbol{f}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\nu_i\right)}{p_{\,\mathrm{tMM}}\left(\boldsymbol{f}_j;\boldsymbol{\Psi}\right)} \tag{15}$$

for $j = 0, \ldots, N_t$ and $i = 1, \ldots, N_f$ and the corresponding posterior probabilities for the latent variables are

$$u_{ij} = \frac{\nu_i + n}{\nu_i + \left(\boldsymbol{f}_j - \boldsymbol{\mu}_i\right)^T\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{f}_j - \boldsymbol{\mu}_i\right)}. \tag{16}$$

In the M-step the parameters in $\boldsymbol{\Psi}$ are recomputed as follows. The component weights are recomputed as the average memberships in each cluster as

$$c_i = \frac{\sum_{j=0}^{N_t}\chi_{ij}}{N_t} \tag{17}$$

whereas the component means are recomputed as the sample averages weighted by the product of posterior probabilities as

$$\boldsymbol{\mu}_i = \frac{\sum_{j=0}^{N_t}\chi_{ij}u_{ij}\boldsymbol{f}_j}{\sum_{j=0}^{N_t}\chi_{ij}u_{ij}} \tag{18}$$

and the component covariances are the average centered outer products weighted by the membership posterior probabilities as

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=0}^{N_t}\chi_{ij}u_{ij}\left(\boldsymbol{f}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{f}_j - \boldsymbol{\mu}_i\right)^T}{\sum_{j=0}^{N_t}\chi_{ij}} \tag{19}$$

for $i = 1, 2, \ldots, N_f$. And the degrees of freedom ν_i is the solution to equation

$$-\psi\left(\frac{1}{2}\nu_i\right) + \frac{1}{\sum_{i=0}^{N_t}\chi_{ij}}\sum_{j=0}^{N_t}\chi_{ij}\left(\ln u_{ij} - u_{ij}\right) + \psi\left(\frac{\nu_i + n}{2}\right) + \ln\left(\frac{1}{2}\nu_i\right) + 1 - \ln\left(\frac{\nu_i + n}{2}\right) = 0, \tag{20}$$

where $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ is the digamma function.

Finally, the formant estimate to join the formant track is

$$F_i = \mu_i \tag{21}$$

for $i = 1, 2, \ldots, N_f$.

TABLE 1. Mean absolute formant estimation errors (Hz) for eight algorithms with vowel /æ/ synthesized in voiced mode from annotations for utterance `w02ae` in Hillenbrand vowel database [19].

Algorithm	F_1	F_2	F_3
hlptmm	39.88	70.29	23.68
hcotmm	39.88	70.29	23.68
hactmm	42.05	73.54	42.74
hlpgmm	38.27	72.35	41.35
stratmm	65.67	179.62	65.50
pyktmm	49.72	73.04	53.36
pykgmm	55.13	113.09	55.36
lpws	42.78	84.41	83.74

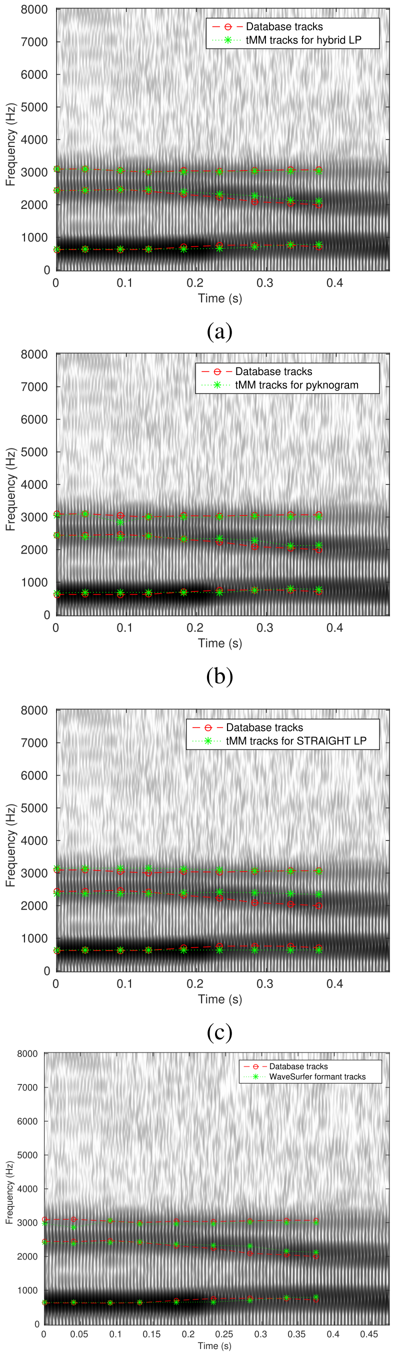
IV. EXPERIMENTAL RESULTS

In the development and test of the resonance frequency estimator and the formant tracker, synthetic and natural speech signals were used. The natural speech utterances are the instances of a set of 12 words with a CVC structure where V is a vowel nucleus, uttered by 50 women, 50 men, 29 boys and 21 girls for a total of 1668 words, which are annotated for the beginning and end times of the vowel nucleus with three formants measured at each whole tens percentage point from 0 through 80% along the length of the vowel nucleus. It is made available by Hillenbrand [18], [19].

For evaluating the impact of voicing in the formant algorithms, two corresponding sets of words were synthesized by an LP formant synthesizer using the annotated data from the Hillenbrand database. Each set of signals is synthesized in a single phonation mode, which is either completely voiced or completely whispered.

The control formant algorithms are pyktmm and pykgmm, where both have formant estimators based on the pyknogram and formant trackers based on tMMs and GMMs [11], respectively, and WaveSurfer (lpws), which is a popular formant algorithm based on linear prediction.[1] The settings for pyktmm are a 100-channel Gabor filterbank with a 20-ms window for the pyknogram, 44.1 kHz upsampling frequency, 4 mixture components, each corresponding to one formant track, and $N_t + 1 = 5$ frames per track step while the settings for lpws are type 0, 12th order LP, 49-ms Hamming windows at 10 ms interval, preemphasis factor 0.7, and 10 kHz downsampling frequency.

Complimentarily, for the selection of the most convenient structure for the proposed formant algorithm, five additional algorithms are evaluated. Four of these algorithms have formant trackers based on tMMs and resonance frequency estimators based on different LP analyzers, namely, hlptmm with the hybrid LP estimator, hcotmm with pitch-synchronous covariance LP estimator, hactmm with the autocorrelation LP estimator, and stratmm with the LP estimator based on the STRAIGHT psd, all as described in Section II. Additionally, another algorithm is hlpgmm, which conjugates the HLP estimator with the GMM formant tracker.



(a)

(b)

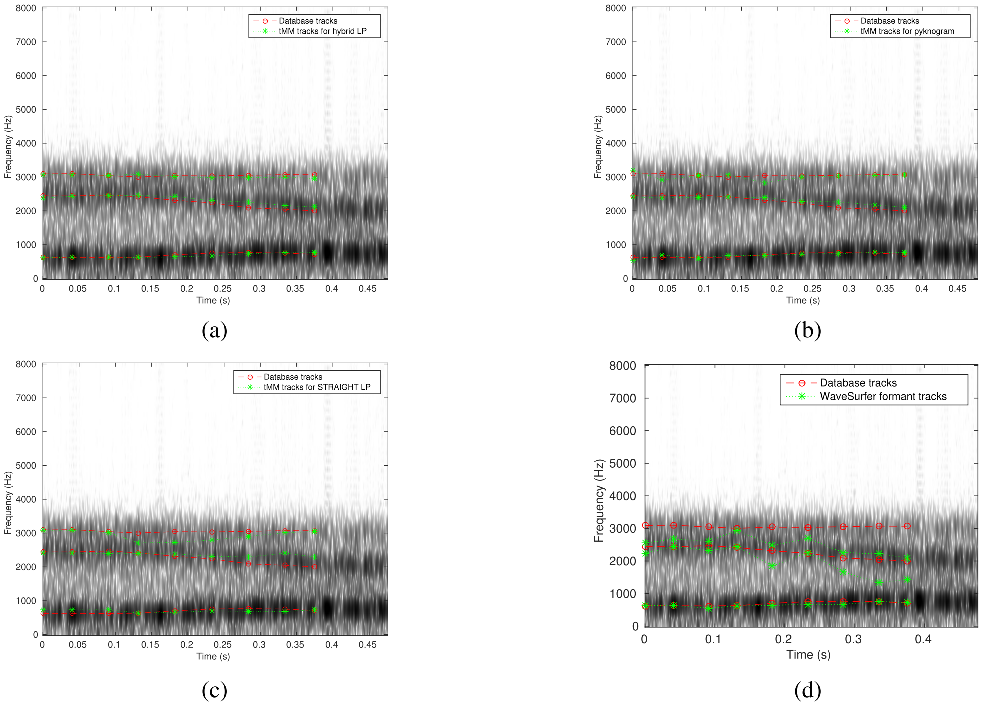
(c)

(d)

FIGURE 4. Estimated formant trajectories overlaid on the spectrogram for the synthetic vowel /æ/ generated in voiced mode from the annotations to female speaker w02 in Hillenbrand vowel database. (a) HLP estimation and tMM tracking. (b) Pyknogram estimation and tMM tracking. (c) STRAIGHT LP estimation and tMM tracking. (d) WaveSurfer estimation and tracking.

The hybrid LP estimator in Section II uses the preemphasis filter $1 - 0.98z^{-1}$ and 20-ms long rectangular windows with 50% overlap for LP order $p = 16$. The estimation of tMMs in Section III uses $N_t + 1 = 6$ frames per track step with N_t frame overlap and univariate densities, that is, $n = 1$.

[1] Software may be downloaded from `https://sourceforge.net/projects/wavesurfer/`.



(a)

(b)

(c)

(d)

FIGURE 5. Estimated formant trajectories overlaid on the spectrogram for the synthetic vowel /æ/ from annotations to female speaker w02 in Hillenbrand vowel database excited in whispering mode. (a) HLP estimation and tMM tracking. (b) Pyknogram estimation and tMM tracking. (c) STRAIGHT LP estimation and tMM tracking. (d) WaveSurfer estimation and tracking.

As a preliminary test, vowel /æ/, as specified for signal `w02ae`, is synthesized in voiced mode and applied to the eight formant algorithms with the resulting mean absolute formant estimation errors shown in Table 1. It is observed that HLP in hlptmm is more accurate than autocorrelation LP and lpws as a vowel formant estimator, reinforcing the principle underlying HLP. As for the tracking method, tMMs perform better than GMMs in vowel formant tracking as can seen by checking the errors for F_3 by hlpgmm and for F_2 by pykgmm. Also, in Fig. 4 d the estimation deviations may be noticed, even though they are relatively small, and in this respect the greatest deviation occurs for the F_2 track of stratmm in Fig. 4 c.

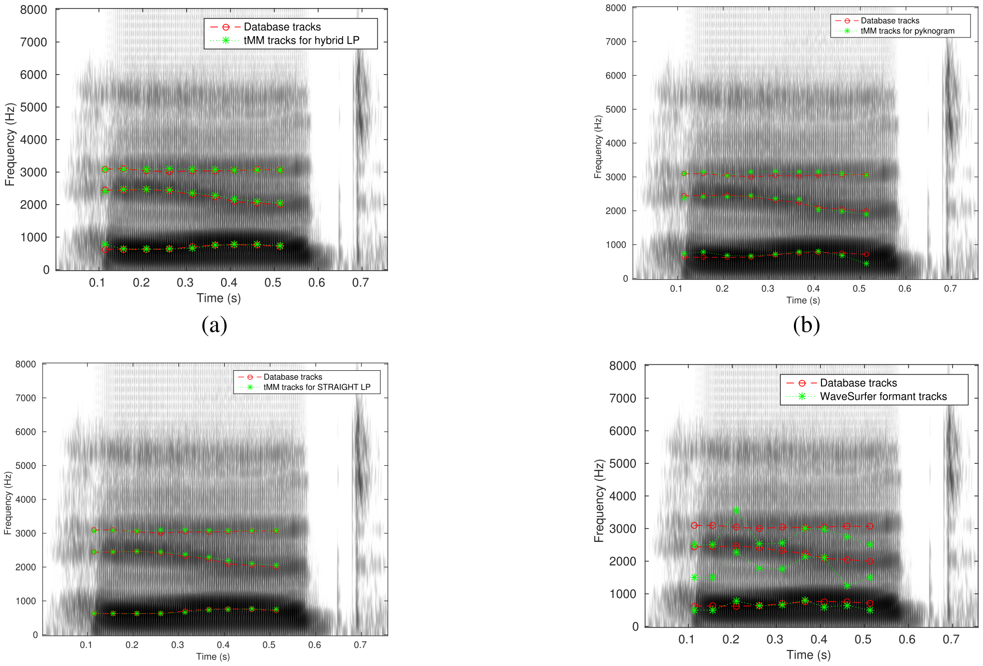
Next the same formant data from speaker w02 is used for synthesizing vowel /æ/ in whispering mode. By comparing the absolute formant estimation errors in Table 2, the choice of autocorrelation LP for unvoiced speech over pitch-synchronous covariance LP analysis is confirmed as preferential. Also, linear prediction performs better than the pyknogram for estimation even though by a smaller margin than for voiced speech. However, stratmm and lpws perform much worse and in the latter case the F_2 and F_3 estimates seem to have been attracted by the actual F_2 track leaving almost untouched the actual F_3 track as observed from Fig. 5d. These may be signs that these frequency estimation methods could have been overly adjusted to voiced speech features. It is still to be noticed that GMM tracking performs comparably to tMM tracking for HLP estimates while it is much worse than tMM tracking in the pyknogram case.

TABLE 2. Mean absolute formant estimation errors (Hz) for eight algorithms with vowel /æ/ synthesized in whispering mode from annotations for utterance `w02ae` in Hillenbrand vowel database [19].

Algorithm	F_1	F_2	F_3
hlptmm	30.80	79.06	57.93
hcotmm	34.68	70.82	75.81
hactmm	30.80	79.06	57.93
hlpgmm	28.13	86.78	48.61
stratmm	59.95	142.34	142.40
pyktmm	45.52	76.33	75.22
pykgmm	41.88	142.84	284.72
lpws	39.96	284.62	553.81

When the whole set of vowel data for all speakers is used, the resulting absolute formant estimation errors are displayed in Table 3 for voiced phonation and in Table 4 for whispered phonation. Hybrid LP is seen to be right in preferring pitch-synchronous covariance analysis to autocorrelation analysis for voiced speech even though by not so great a margin as for the single signal whereas its preference of autocorrelation analysis to pitch-synchronous covariance analysis in the whispering case is surely correct. As for the other LP frequency estimators, both stratmm and lpws perform worse, even though stratmm is not so bad in the voiced case. On the other hand, tMM tracking is superior to GMM either combined with LP estimation or pyknogram estimation in the voiced or the whispered case.

When the natural utterance from speaker w02 containing vowel /æ/ is processed by the formant algorithms, the results are mostly better than those obtained with the synthetic



(a)

(b)

(c)

(d)

FIGURE 6. Estimated formant trajectories overlaid on the spectrogram for the utterance "head", which includes the vowel /æ/ from female speaker w02 in Hillenbrand vowel database. (a) HLP estimation and tMM tracking. (b) Pyknogram estimation and tMM tracking. (c) STRAIGHT LP estimation and tMM tracking. (d) WaveSurfer estimation and tracking.

TABLE 3. Mean absolute formant estimation errors (Hz) for eight algorithms with vowels synthesized in voiced mode from Hillenbrand men, women and children vowel database annotations [19].

Algorithm	F_1	F_2	F_3
hlptmm	40.84	115.46	322.01
hcotmm	40.84	115.46	322.01
hactmm	36.43	95.85	414.80
hlpgmm	53.08	175.01	552.28
stratmm	49.63	162.26	342.59
pyktmm	108.98	357.07	334.88
pykgmm	125.11	416.02	423.38
lpws	34.47	330.73	337.90

TABLE 4. Mean absolute formant estimation errors (Hz) for eight algorithms with vowels synthesized in whispering mode from Hillenbrand men, women and children vowel database annotations [19].

Algorithm	F_1	F_2	F_3
hlptmm	28.22	109.47	385.75
hcotmm	42.25	194.38	328.59
hactmm	28.22	109.47	385.75
hlpgmm	56.87	176.87	489.96
stratmm	70.75	206.29	402.98
pyktmm	110.15	350.69	353.73
pykgmm	145.52	463.75	488.90
lpws	31.74	203.24	657.75

signals as shown in Table 5, where the LP estimation methods are very nearly comparable with the exception of lpws, which displays the same tracking problems already observed with the synthetic whispered version of the signal as can be seen by referring to Fig. 6 d. As a minor event in comparison, a more pronounced deviation from the F_1 track may be observed for the estimated pyktmm track in Fig. 6 b. Still from Table 5, clustering by tMMs is seen to perform much better than clustering by GMMs, especially in the HLP case.

TABLE 5. Mean absolute formant estimation errors (Hz) for eight algorithms with natural vowel /æ/ from utterance `w02ae` in Hillenbrand vowel database [19].

Algorithm	F_1	F_2	F_3
hlptmm	27.45	38.97	30.61
hcotmm	25.66	37.16	35.46
hactmm	26.89	34.14	32.61
hlpgmm	1341.89	511.26	654.76
stratmm	13.91	42.51	36.73
pyktmm	86.48	61.19	66.58
pykgmm	108.34	86.82	228.09
lpws	107.46	523.03	407.34

TABLE 6. Mean absolute formant estimation errors (Hz) for eight algorithms with natural vowels from Hillenbrand men, women and children vowel database [19].

Algorithm	F_1	F_2	F_3
hlptmm	59.83	139.99	350.59
hcotmm	61.30	156.02	347.09
hactmm	56.62	139.88	363.43
hlpgmm	1263.88	808.45	765.85
stratmm	61.46	133.16	345.41
pyktmm	148.81	388.07	393.85
pykgmm	231.25	356.38	443.20
lpws	145.16	606.36	744.87

When the formant algorithms are applied to the complete set of natural utterances in the Hillenbrand database, the absolute deviations from annotated values are mostly comparable to those of the voiced synthetic case as can be seen by referring to Table 6 in comparison to Table 3. In Table 6 the performance of hlptmm conjugates the better performance of hcotmm for F_3 with the better performance of hactmm for the lower formants. Further, the performance of hlptmm is comparable to stratmm and it is superior to lpws for every formant frequency. In addition, the performance of tMM clustering is superior to that of GMM clustering in both the LP and the pyknogram estimation cases, even more so for the former, thus underlining the fact that the t density better fits the distribution of frequency estimates than the Gaussian density.

V. CONCLUSION

A formant tracker was proposed consisting of a novel hybrid autoregressive vocal tract resonance frequency estimator followed by a tMM tracking algorithm. The tMM tracker was proposed previously in connection with a pyknogram time-frequency representation. The hybrid autoregressive estimator consists of an accurate integer-cycle pitch-synchronous covariance LP analysis for voiced speech and autocorrelation analysis for unvoiced speech. Evidence and considerations into the nature of vocal tract resonances are presented to justify that tMMs fit formant trajectories more closely than GMMs. In tests with voiced and whispered synthetic speech and with an annotated database, the proposed formant tracker is shown to provide closer or comparable estimates in comparison with the pyknogram tMM tracker and other LP estimators, including the popular WaveSurfer LP tracker and an LP estimator based on the STRAIGHT power spectral density and also clustered by tMMs, so that the improved performance results from the harmonious combination of hybrid linear prediction for resonance estimation and tMM for formant tracking.

ACKNOWLEDGMENT

The author is grateful to Dr. Harshavardhan Sundar, Prof. Chandra Sekhar Seelamantula and Prof. Thippur V. Sreenivas for sharing the implementation of their formant tracker [11].

REFERENCES

[1] I. R. Titze *et al.*, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *J. Acoust. Soc. Amer.*, vol. 137, no. 5, pp. 3005–3007, May 2015.

[2] L. Deng, L. J. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, vol. 1, May 2004, pp. 557–560.

[3] D. T. Toledano, J. G. Villardebó, and L. H. Gómez, "Initialization, training, and context-dependency in HMM-based formant tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 511–523, Mar. 2006.

[4] E. Özkan, İ. Y. Özbek, and M. Demirekler, "Dynamic speech spectrum representation and tracking variable number of vocal tract resonance frequencies with time-varying Dirichlet process mixture models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1518–1532, Nov. 2009.

[5] K. Xia and C. Y. Espy-Wilson, "A new strategy of formant tracking based on dynamic programming," in *Proc. ICSLP*, Beijing, China, vol. 3, 2000, pp. 55–58.

[6] T. Kaneko and T. Shimamura, "Noise-reduced complex LPC analysis for formant estimation of noisy speech," *Int. J. Electron. Elect. Eng.*, vol. 2, no. 2, pp. 90–94, Jun. 2014.

[7] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and Bayesian estimation for robust formant tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 224–236, Feb. 2010.

[8] T. Kinnunen, B. Zhang, J. Zhu, and Y. Wang, "Speaker verification with adaptive spectral subband centroids," in *Advances in Biometrics* (Lecture Notes in Computer Science), vol. 4642, S.-W. Lee and S. Li, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 58–66.

[9] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, USA, vol. 2, May 1998, pp. 617–620.

[10] H. Oohashi, S. Hiroya, and T. Mochida, "Real-time robust formant tracking system using a phase equalization-based autoregressive exogenous model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5118–5121.

[11] H. Sundar, C. S. Seelamantula, and T. V. Sreenivas, "A mixture model approach for formant tracking and the robustness of Student's-t distribution," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2626–2636, Dec. 2012.

[12] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar./Apr. 2008, pp. 3933–3936.

[13] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 5535–5539.

[14] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Amer.*, vol. 134, no. 2, pp. 1295–1313, Aug. 2013.

[15] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Germany: Springer, 1976.

[16] M. A. Ramírez, "A Levinson algorithm based on an isometric transformation of Durbin's," *IEEE Signal Process. Lett.*, vol. 15, pp. 99–102, Jan. 2008.

[17] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York, NY, USA: Wiley, 2000.

[18] J. M. Hillenbrand. Vowel Database. Western Michigan University. Accessed: Jun. 27, 2016. [Online]. Available: https://homepages.wmich.edu/~hillenbr/voweldata.html

[19] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Amer.*, vol. 97, no. 5, pp. 3099–3111, May 1995.

MIGUEL ARJONA RAMÍREZ (M'78–SM'00) received the B.S. degree in electronics engineering from the Instituto Tecnológico de Aeronáutica, Brazil, in 1980, the graduate degree in electronic design engineering from the Philips International Institute, The Netherlands, in 1981, and the M.S. and Ph.D. degrees in electrical engineering and Habilitation in signal processing from the University of São Paulo, Brazil, in 1992, 1997, and 2006, respectively.

He was an Engineering Development Group Leader for interactive voice response systems with Itautec Informática, Brazil, where he served from 1982 to 1990. In 2008, he carried research in time-frequency speech analysis and coding in a research visit to the Royal Institute of Technology in Sweden. He is currently an Associate Professor with Escola Politécnica, University of São Paulo, where he is also a member of the Signal Processing Laboratory. He has authored or co-authored four book chapters and over 60 journal and conference papers in these areas. His research focuses on the application of novel signal processing and machine learning algorithms to signal compression and prediction, speech analysis, coding and recognition, speaker identification, and audio analysis and coding. He is a member of the Brazilian Telecommunications Society.