NOVEL AUDITORY MOTIVATED SUBBAND TEMPORAL ENVELOPE BASED FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHM

S Chandra Sekhar [1], *Sridhar Pilli* [2], *Lakshmikanth C* [3] *and TV Sreenivas* [4]

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore - 560 012, India.
[1]schash,[4]tvsree@ece.iisc.ernet.in,[2]sridhar.pilli,[3]lakshmikanth.c@gmail.com

ABSTRACT

We address the problem of estimating the fundamental frequency of voiced speech. We present a novel solution motivated by the importance of amplitude modulation in sound processing and speech perception. The new algorithm is based on a cumulative spectrum computed from the temporal envelope of various subbands. We provide theoretical analysis to derive the new pitch estimator based on the temporal envelope of the bandpass speech signal. We report extensive experimental performance for synthetic as well as natural vowels for both real-world noisy and noise-free data. Experimental results show that the new technique performs accurate pitch estimation and is robust to noise. We also show that the technique is superior to the autocorrelation technique for pitch estimation.

1. INTRODUCTION

The problem of estimating the fundamental frequency of voicing [1], F_0, is important in many speech applications, such as compression, speech synthesis, enhancement in the presence of noise etc. Precise estimates of F_0 are very useful in natural sounding synthesized speech and speech signal compression for efficient transmission. Also, F_0 tracking is an important technique in the analysis of stressed speech since it is associated with various factors such as emotions, change in muscle activity, blood pressure, heart rate [2] etc. Pitch estimation and tracking is also useful in extracting melody patterns in music signals, which are useful in *query-by-humming* audio retrieval systems [3]. Several algorithms have been proposed in the past for the problem of pitch estimation [1, 4] and newer techniques or modifications to existing techniques continue to appear in the literature indicating that the problem is still of immense research interest [5, 6]. In this paper, we propose a new technique motivated by the auditory peripheral processing of the acoustic stimuli.

We briefly review the auditory processing of acoustic stimuli. The current models for the inner ear [7, 8] describe the mechanical motion at every point along the basilar membrane as the output of a bandpass filter with frequency response determined by the mechanical tuning characteristics at that location. The shearing movement between the tectorial membrane and the basilar membrane causes the inner hair cell cilia to bend resulting in an electric discharge in the auditory nerve fibres, in a nonlinear manner. The nerve fibres, broadly characterized into medium, low and high rate fibres, are characterized by a threshold level and spontaneous rate of firing. The instantaneous discharge rate of auditory nerve fibres is found to be maximum during the initial 15ms of the acoustic stimulation and then decreasing, until it reaches a steady-state, about 50ms after the stimulus onset. The decrease in the fibre response rate is the result of adaptation to the temporal envelope of the subband output. In addition to the temporal envelope dynamics of the response, the response also includes the detailed timing behaviour of the response to each input cycle [7]. The auditory nerve fibres tend to fire in a phase-locked manner to low-frequency periodic stimuli. However, the fibres responsive to the high-frequency components of a signal tend to synchronize only to the amplitude modulation/envelope of the signal [7]. The envelope-synchronized nerve firing happens at the signal's fundamental frequency. The envelope is also a measure of the average nerve fibre discharge rate response in that channel.

A set of research results on the importance of the subband temporal envelope processing mechanism [9, 10] for improved speech perception strongly supports the existence of auditory pathways, exclusive and highly specialized to process subband signal amplitude modulation. It is not clearly known if these modulation-specific auditory pathways also capture the pitch information apart from improving speech perception.

Another recent research on the neuronal representation of pitch and pitch perception [11] has shown the existence of neurons in the auditory cortex of marmoset monkeys that respond to both pure tones and missing fundamental harmonic complex sounds with the same fundamental frequency providing a neural correlate for pitch constancy. Perhaps, it is due to this feature that the pitch is perceived even in the case of sounds with missing fundamental such as those encountered in speech transmission over telephone channel.

Our research reported in this paper is motivated by these results on auditory sound processing and pitch perception and is strongly supported by some interesting observations on the temporal properties of the bandpass signals of periodic/voiced sounds. We show that the signal's fundamental frequency can be estimated from the temporal envelope of the bandpass signal. We develop the new technique for fundamental frequency estimation and study its accuracy by considering synthetic voiced data. We also show its robustness to noise by considering synthetic and natural voiced sounds in the presence of various kinds of noise and different signal to noise ratio (SNR).

2. TEMPORAL ENVELOPE OF BANDPASS SPEECH SIGNAL

In the source-filter model for speech production, the voiced sounds are modeled as the output of a multi-pole autoregressive system driven by a quasi-periodic excitation. Let $g(t)$ denote the quasi-periodic excitation. and $v(t)$ denote the vocal tract impulse response. The output of the vocal tract is given by $s(t) = v(t) * g(t)$. In order to avoid mathematical complication, we assume that $g(t)$ is periodic and write $g(t) = \sum_k p(t + kT_0)$ where T_0 is the pitch period and $p(t)$ is the fundamental period of $g(t)$.

2.1 At resonance

Consider the speech signal $s(t)$, bandpass-filtered around a formant to yield $s_k(t)$. We can write $s_k(t) = s(t) * h_k(t)$, where $h_k(t)$ is the impulse response of a filter centered about a formant. The filter could be one of the filters in a typical auditory filterbank. The filter output is periodic and given by

$$\begin{aligned} s_k(t) &= \left(v(t) * \sum_k p(t + kT_0)\right) * h_k(t) \\ &= \Big(v(t) * h_k(t)\Big) * \left(\sum_k p(t + kT_0)\right). \end{aligned} \quad (1)$$

$v(t) * h_k(t)$ is a decaying sinusoid of the form $e^{-\alpha_k t}\sin(\omega_k t + \phi_k)u(t)$ where $u(t)$ is the unit-step function, α_k is the damping factor and determines the formant bandwidth and ω_k is the formant frequency in radians. In writing so, we have assumed that the effect of the neighbouring formants is negligible. Therefore,

$$s_k(t) \approx e^{-\alpha_k t}\sin(\omega_k t + \phi_k)u(t) * \left(\sum_k p(t + kT_0)\right). \quad (2)$$

In the impulse train excitation model, $p(t) = \delta(t)$,

$$s_k(t) = \sum_m e^{-\alpha_k(t+mT_0)}\sin(\omega_k(t+mT_0)+\phi_k)u(t+mT_0). \quad (3)$$

Without loss of generality, assume $\phi_k = 0$. To enable further analysis, let us consider the complex form of $s_k(t)$, given by,

$$a_{s_k}(t) = e^{-(\alpha_k + j\omega_k)t}\sum_m e^{-(\alpha_k + j\omega_k)mT_0}u(t + mT_0). \quad (4)$$

The imaginary part of $a_{s_k}(t)$ is $s_k(t)$. If the damping factor α_k, is such that the leakage of the vocal tract impulse response into neighbouring pitch periods is negligible, then, over a pitch period, only one term in the summation is significant. Therefore, over $[0, T_0]$,

$$a_{s_k}(t) \approx e^{-(\alpha_k + j\omega_k)t)}[u(t) - u(t - T_0)]. \quad (5)$$

corresponding to a pitch pulse at an arbitrary position $t = 0$. In general, over $[mT_0, (m+1)T_0]$, we can write,

$$\begin{aligned} a_{s_k}(t) &= e^{-(\alpha_k + j\omega_k)t}e^{-(\alpha_k + j\omega_k)mT_0} \\ &\quad \big(u(mT_0) - u(t - (m+1)T_0)\big). \end{aligned} \quad (6)$$

For $t \in [mT_0, (m+1)T_0]$, the envelope is given by

$$e^{-\alpha_k(t+mT_0)}\big[u(t - mT_0) - u(t - (m+1)T_0)\big]. \quad (7)$$

Thus, the envelope can be written approximately as,

$$e_k(t) \approx \sum_m e^{-\alpha_k(t+mT_0)}\big[u(t - mT_0) - u(t - (m+1)T_0)\big], \quad (8)$$

which is periodic with period T_0. We can rewrite $e_k(t)$ as:

$$e_k(t) = \tilde{e}^{-\alpha_k t} * \sum_m \delta(t + mT_0), \quad (9)$$

where

$$\tilde{e}^{-\alpha_k t} = e^{-\alpha_k t}\big(u(t) - u(t - T_0)\big). \quad (10)$$

The spectrum of $e_k(t)$ is given by

$$E_k(\omega) = \tilde{E}_k(\omega)\frac{1}{T_0}\sum_m \delta\left(\omega - \frac{2\pi}{T_0}m\right), \quad (11)$$

where $\tilde{E}_k(\omega) = \mathcal{F}(\tilde{e}^{-\alpha_k t})$, where $\mathcal{F}(.)$ denotes the Fourier transform.

Consider $e_k(t)$ windowed by $w(t)$, over a certain duration $[0, T_1]$. The spectrum of $e_k(t)w(t)$ is given by,

$$\tilde{E}(\omega)\frac{1}{T_0}\sum_m W\left(\omega - \frac{2\pi}{T_0}m\right), \quad (12)$$

where $W(\omega) = \mathcal{F}(w(t))$. To reduce the effect of the term $\tilde{e}^{-\alpha_k t}$, we perform differentiation of the envelope, $w(t)e_k(t)$ to yield

$$\frac{d[w(t)e_k(t)]}{dt} = \frac{d}{dt}\Big[\big(\tilde{e}^{-\alpha_k t} * \delta(t + mT_0)\big)w(t)\Big]. \quad (13)$$

The spectrum of $\frac{d[w(t)e_k(t)]}{dt}$ is given by,

$$j\omega\tilde{E}(\omega)\frac{1}{T_0}\sum_m W\left(\omega - \frac{2\pi}{T_0}m\right). \quad (14)$$

The peak location of the spectrum is an estimate of the fundamental frequency, $\frac{1}{T_0}$. We use the Hilbert transform technique for computing the temporal envelope. In discrete-time implementation, the differentiation operation is replaced by the difference operator with Z-transform: $1 - z^{-1}$. This boosts the high frequency component energy and compensates for the decaying frequency response $\tilde{E}_k(\omega)$, resulting in a dominant peak corresponding to the fundamental frequency.

2.2 Off resonance

Consider the bandpass spectrum in regions that are away from formant resonances. Performing approximate analysis, we show that the envelope is related to the fundamental frequency. Let the output be given by

$$\begin{aligned} s_k(t) &= A_1\cos(m\omega_0 t) + A_2\cos((m+1)\omega_0 t) \\ &\quad + A_3\cos((m+2)\omega_0 t), \end{aligned} \quad (15)$$

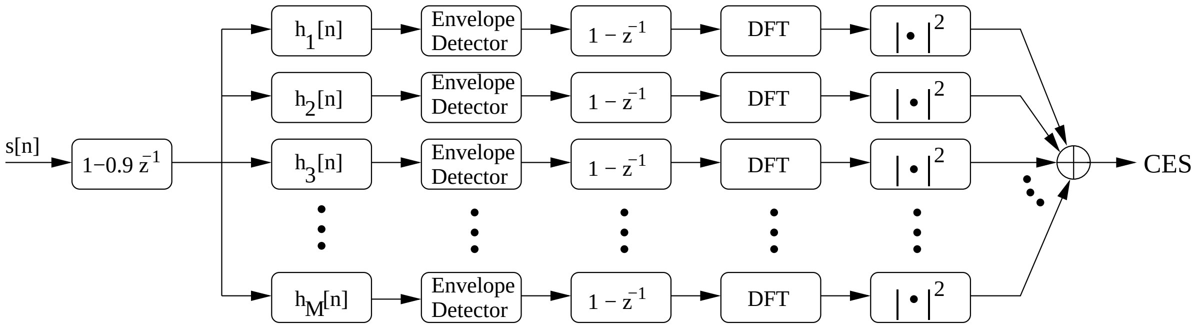

Figure 1: Cumulative Envelope Spectrum (CES) computation.

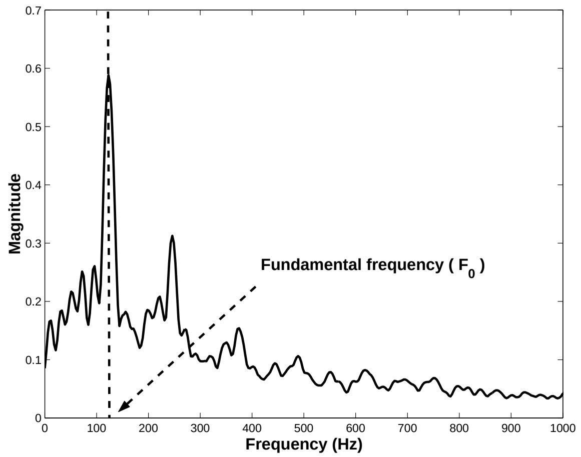

Figure 2: Cumulative envelope spectrum for a natural vowel.

i.e., the harmonic frequencies present are $m\omega_0, (m+1)\omega_0, (m+2)\omega_0$ rad/s where $\omega_0 = \frac{2\pi}{T_0}$. The amplitudes A_1, A_2 and A_3 depend on the filter responses as well as the corresponding input strengths. Given the frequency-selective nature of the auditory filters, we can write, $A_2 \approx 2A_1 \approx 2A_3$, i.e., we have assumed that the spectrum component at $(m+1)\omega_0$ is nearly twice as dominant as the components at $m\omega_0$ and $(m+2)\omega_0$. This simplifying assumption and gives good insight into the envelope relation with the pitch period. The output can then be simplified to yield, $s_k(t) \approx 2\cos((m+1)\omega_0 t)\,(1+\cos(\omega_0 t))$. The envelope is given by $e_k(t) = 2(1+\cos(\omega_0 t))$. Its derivative is $\frac{de_k(t)}{dt} = -2\omega_0 \sin(\omega_0 t)$, which is periodic and its spectrum has a peak at the frequency ω_0.

3. CUMULATIVE ENVELOPE SPECTRUM

In the above analysis, we have shown that the subband temporal envelopes, at or off-formant resonance, are periodic with the same period as the pitch period and hence can be estimated by a periodicity analysis of the temporal envelope. To avoid the problem of formant frequency estimation, which is not known apriori [5], we take an auditory filterbank based frontend processing approach for F_0 estimation. We approximate the peripheral auditory filterbank analysis system by 19 bandpass filters, which collectively cover the frequency range from 100Hz to 4000Hz. The bandwidth of the channels is half-Bark, where a Bark is the critical bandwidth. The filters are implemented using the Malcolm Slaney's Auditory Toolbox [12]. $w(t)e_k(t)$ is the envelope of the k^{th} subband signal and $E_k(\omega) = \mathcal{F}\big(\frac{d}{dt}w(t)e_k(t)\big)$. We used the rectangular window function $w(t)$.

Vowel	*Clean*	*10dB*		*20dB*	
		μ	σ	μ	σ
IY	150.4	162.9	29.4	151.2	0.8
IH	150.0	150.9	2.1	150.3	0.6
EH	149.5	149.0	1.8	148.7	0.6
AE	149.4	148.6	1.4	148.7	0.4
AH	149.5	150.9	11.5	148.8	1.1
AA	149.2	148.7	6.8	147.1	2.5
AO	149.2	154.2	17.9	147.2	2.3
UH	149.9	165.7	33.8	148.0	2.8
UW	148.8	196.5	38.9	150.3	17.5
ER	149.4	149.8	4.9	148.9	0.9

Table 1: F_0 estimation performance using the CES algorithm for synthetic vowel data.

The subband envelope spectra are combined to yield the *cumulative envelope spectrum* as:

$$\mathcal{E}(\omega) = \sum_{k=1}^{M} |E_k(\omega)|^2, \qquad (16)$$

where M is the number of equivalent rectangular bandwidth (ERB) filters. In practice, we have sampled-data and hence we use the discrete Fourier transform. The sequence of operations is shown in the form of a block diagram in Fig. 1. The pre-emphasis filter has a Z-transfer function given by $1 - 0.95z^{-1}$. This approximately models the broad outer-ear resonances causing a 10-20dB boost in energy between 1.5-5kHz. As a result, the spectrum dynamic range is reduced and the subband temporal envelopes in this frequency range get a higher weightage relative to those at the output of the filters with lower center frequency. The cumulative envelope spectrum is actually a time-varying spectrum since it is computed over a short duration. Therefore, we represent it as $\mathcal{E}[\ell, m]$ where ℓ refers to the frame index and m refers to the DFT index. The DFT can be efficiently implemented using the FFT algorithm. Alternatively,

	Clean		10dB				20dB			
	CES	AUTO	CES		AUTO		CES		AUTO	
Vowel	F_0	F_0	μ	σ	μ	σ	μ	σ	μ	σ
AAnderstand (understand, female)	171.6	168.5	171.8	4.3	144.7	40.8	171.5	0.7	168.9	0.9
attEHined (attained, male)	135.9	129.4	136.4	0.9	123.5	20.3	136.1	0.2	129.5	0.4
portugIYEse (portugese, male)	123.4	121.5	128.2	23.4	119.2	33.7	123.4	0.1	121.4	0.7
AApon (upon, male)	105.4	106.4	104.7	1.1	99.1	23.8	105.4	0.3	103.9	12.1
cEIves (caves, female)	165.3	157.4	173.7	19.1	151.7	22.2	165.4	0.7	157.5	0.6
gUHd (good, female)	155.5	148.1	170.4	20.2	138.9	28.3	155.2	1.5	148.3	0.3
cAEpitalised (capitalised, male)	131.2	129.6	131.1	1.1	120.4	37.6	131.2	0.3	125.5	16.7
ERned (earned, male)	127.9	130.3	127.3	1.6	119.9	25.9	127.8	0.5	126.9	17.5
sUWn (soon, female)	143.2	142.2	143.4	1.2	141.9	0.8	143.2	0.3	141.9	0.4
dAHtch (dutch, female)	139.1	136.0	139.5	1.7	129.5	58.1	139.1	0.3	133.3	13.5
britIHsh (british, male)	120.8	119.8	123.1	17.4	118.4	37.3	120.9	0.2	119.8	0.3
pAO (paw, female)	132.4	133.0	134.9	14.6	124.8	40.9	131.2	1.3	132.6	6.7
lIHved (lived, female)	165.3	161.8	173.7	17.4	143.6	34.5	165.1	2.4	161.8	0.4

Table 2: F_0 estimation performance of the CES and autocorrelation algorithms for natural data in **cordless phone channel noise**. The vowel portion is indicated in uppercase. The gender of the speaker and the word are also indicated.

	Clean		10dB				20dB			
	CES	AUTO	CES		AUTO		CES		AUTO	
Vowel	F_0	F_0	μ	σ	μ	σ	μ	σ	μ	σ
AAnderstand (understand, female)	171.6	168.5	171.5	3.6	154.5	33.5	171.5	0.6	168.9	0.7
attEHined (attained, male)	135.9	129.4	136.3	0.9	126.3	15.3	136.1	0.2	129.5	0.4
portugIYEse (portugese, male)	123.4	121.5	126.9	20.4	127.9	38.8	123.4	0.1	121.4	0.7
AApon (upon, male)	105.4	106.4	104.7	1.1	103.8	37.2	105.4	0.3	104.3	10.8
cEIves (caves, female)	165.3	157.4	170.3	17.6	151.0	22.8	165.4	0.6	157.5	0.5
gUHd (good, female)	155.5	148.1	164.1	17.9	140.2	26.8	155.3	1.4	148.2	0.3
cAEpitalised (capitalised, male)	131.2	129.6	131.0	0.9	124.0	28.7	131.1	0.3	127.7	11.9
ERned (earned, male)	127.9	130.3	127.3	1.7	120.5	26.7	127.8	0.5	128.4	13.9
sUWn (soon, female)	143.2	142.2	143.4	1.2	141.9	0.8	143.2	0.3	141.9	0.4
dAHtch (dutch, female)	139.1	136.0	139.5	1.5	137.3	58.3	139.1	0.3	134.9	9.7
britIHsh (british, male)	120.8	119.8	123.1	17.2	117.8	28.3	120.8	0.2	119.8	0.3
pAO (paw, female)	132.4	133.0	133.0	11.1	127.5	35.9	131.2	1.3	132.9	4.7
lIHved (lived, female)	165.3	161.8	169.1	15.2	147.1	32.6	169.9	1.9	161.8	0.4

Table 3: F_0 estimation performance of the CES and autocorrelation algorithms for natural data in **mobile phone channel noise**. The vowel portion is indicated in uppercase. The gender of the speaker and the word are also indicated.

	Clean		10dB				20dB			
	CES	AUTO	CES		AUTO		CES		AUTO	
Vowel	F_0	F_0	μ	σ	μ	σ	μ	σ	μ	σ
AAnderstand (understand, female)	171.6	168.5	171.2	3.0	153.5	34.9	171.5	0.6	169.0	0.8
attEHined (attained, male)	135.9	129.4	136.3	0.8	125.4	16.7	136.1	0.2	129.5	0.4
portugIYEse (portugese, male)	123.4	121.5	127.6	21.9	128.9	42.9	123.4	0.1	121.3	0.6
AApon (upon, male)	105.4	106.4	104.8	1.1	107.5	41.9	105.4	0.3	104.3	10.4
cEIves (caves, female)	165.3	157.4	169.1	14.8	145.9	30.8	165.4	0.6	157.5	0.5
gUHd (good, female)	155.5	148.1	160.5	16.0	139.7	27.8	155.4	1.3	147.9	4.9
cAEpitalised (capitalised, male)	131.2	129.6	131.1	0.9	121.1	28.9	131.2	0.3	127.4	12.8
ERned (earned, male)	127.9	130.3	127.2	1.6	119.9	27.3	127.8	0.5	128.6	13.3
sUWn (soon, female)	143.2	142.2	143.3	1.1	141.9	0.8	143.2	0.3	141.9	0.4
dAHtch (dutch, female)	139.1	136.0	139.6	1.5	136.6	60.6	139.1	0.3	135.1	15.3
britIHsh (british, male)	120.8	119.8	122.5	14.9	116.2	26.9	120.9	0.2	119.8	0.3
pAO (paw, female)	132.4	133.0	131.9	8.9	123.9	38.3	131.2	1.2	133.0	4.7
lIHved (lived, female)	165.3	161.8	167.7	15.1	147.1	32.9	164.9	1.7	161.9	0.4

Table 4: F_0 estimation performance of the CES and autocorrelation algorithms for natural data in **landline telephone channel noise**. The vowel portion is indicated in uppercase. The gender of the speaker and the word are also indicated.

the Goertzel's algorithm can be used since we need not compute the spectrum from 0 to $\frac{F_s}{2}$Hz. In our simulations we restrict the DFT computation to [90 , 250] Hz spectrum region. This can be varied depending on any apriori information about the speaker population.

The pitch period for the ℓ^{th} frame is estimated as: $\hat{T}_0(\ell) = \frac{1}{\hat{F}_0(\ell)}$. where $\hat{F}_0(\ell) = arg \max_m \ \mathcal{E}[\ell, m]$. To reduce the errors due to sampling, we perform quadratic-curve fitting about the sampled spectrum maximum to yield a better estimate of F_0. The cumulative envelope spectrum for a natural vowel is shown in Fig. 2. Note the well-defined peak at F_0.

4. EXPERIMENTAL RESULTS

To assess the accuracy of the CES pitch estimation algorithm, we perform experiments on synthetic as well as natural data. The synthetic vowel data is generated using the auditory toolbox, at a sampling frequency of 8kHz, with the fundamental frequency chosen as 150Hz. We use 20ms window for processing. The results of pitch estimation using the CES algorithm are shown in Table. 1. To assess the robustness of the CES pitch estimator, we add noise of a required variance to generate a noisy signal (generated using a pseudorandom, white Gaussian noise generator) of a desired signal-to-noise ratio (SNR). Based on 100 such realizations, we obtain the mean (μ) and the standard deviation (σ) of the F_0 estimate (in Hz). The results are shown in Table. 1 for two different values of the SNR. We observe that the new CES pitch estimator is robust to noise. The average values of the fundamental frequency are quite close to the actual pitch value (150Hz) even at 10dB global SNR. The synthetic data performance analysis is to enable a comparison to the actual value of the pitch. This is not possible with natural data. The vowel for which the algorithm showed consistent performance is /AE/. The vowel for which poor performance was obtained is /UW/.

The performance of the algorithm for natural vowels in real-world noises is investigated next. We used the Carnegie Mellon University's database [13] for the experiments . We manually sliced out portions containing vowel data in the natural speech recording. We used the cordless phone, mobile phone and landline phone noise data from the Indian Institute of Science - BPL database (noisy speech database created in the Speech and Audio Lab, Indian Institute of Science). The mean and standard deviation in fundamental frequency estimates, for natural data, are shown in Tables. 2,3 and 4. The pitch estimates for noise-free data are also shown for the purpose of comparison. The results are obtained from 100 realizations of the noisy data. The results show that the new technique (CES) is quite robust to noise and superior to the autocorrelation (AUTO) technique for F_0 estimation.

Ideally, it is desired that F_0 estimation be independent of the vowel formant frequencies. However, the variation of the F_0 estimation accuracy from one vowel to another (synthetic and natural) shows that the pitch estimation algorithm is dependent on the formant frequencies. The exact nature of the dependency requires further investigation.

5. CONCLUSIONS

Motivated by the importance of amplitude modulation in temporal processing of speech signals, we presented a novel F_0 estimation algorithm. Using synthetic data, we showed that the technique is very accurate. We also demonstrated that the technique is robust to noise and superior in performance to the autocorrelation technique for both synthetic and real-world noisy speech data. We can also modify the technique to improve its performance. For example, weighing the subband envelope spectrum by the subband SNR increases noise-robustness. The cumulative envelope spectrum technique is also suitable for application in speech coding and query-by-humming audio retrieval systems. These results will be reported separately.

REFERENCES

[1] W. Hess, "Pitch determination of speech signals", *Springer Verlag*, 1983.

[2] K. Gopalan, "Pitch estimation using a modulation model of speech", *Proc.* 5^{th} *Intl. Conf. on Sig. Proc.*, WCCC-ICSP, Vol. 2, pp. 786-791, Aug 2000.

[3] R.J.McNab, L.A. Smith, I.H. Witten, C.L. Henderson, S.J. Cunningham, "Towards the digital music library: Tune retrieval from acoustic input", *Proc. Digital Libraries*, pp 11-18, 1996.

[4] T.V. Sreenivas and P.V.S. Rao, 'Pitch extraction from corrupted harmonics of the power spectrum', *J. Acoust. Soc. of America*, Vol.65, No.1, pp 223-228, Jan 1979.

[5] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch estimation of noisy speech", *IEEE Trans. Speech and Audio Proc.*, Vol. 9, No. 7, pp. 727-730, Oct. 2001.

[6] K. Kasi and S.A. Zahorian, "Yet another algorithm for pitch tracking", *Proc. IEEE Intl. Conf. on Acoust. Speech and Sig. Proc.* (ICASSP), Vol. 1, pp. 361-364, May 2002.

[7] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing", *Jl. of Phonetics*, Vol. 16, pp. 55-76, 1988.

[8] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition", *Jl. Acoust. Soc. Am.* 106(4), pp. 2040-2050, Oct. 1999.

[9] R.V. Shannon, "Evidence from auditory brainstem implants of a modulation-specific auditory pathway that is critical for speech recognition", *DeVault Lab Colloquium*, 15 April 2005, Indiana University.

[10] Q.J-Fu, "Temporal processing and speech recognition in cochlear implant users", *Neuro Report*, pp. 1-5, vol. 13, no. 13, Sep 2002.

[11] D. Bendor and X. Wang, "The neuronal representation of pitch in primate auditory cortex", *Letters, Nature*, Vol. 436, pp. 1161-1165, 25 Aug 2005.

[12] M. Slaney, "Auditory Toolbox", Version 2, Source: http://rvl4.ecn.purdue.edu/ malcolm/interval/1998-010/

[13] http://www.festvox.org/cmu_arctic