ROBUST VOICING ESTIMATION WITH DYNAMIC TIME WARPING

Tian Wang and Vladimir Cuperman

Dept. of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106

ABSTRACT

This paper[*] presents a robust voicing estimation algorithm for low bit rate harmonic speech coding. The algorithm is based on waveform time-warping followed by spectral matching based on voiced and unvoiced local spectral models. The objective of time warping is to reduce the effect of pitch variations on the voicing decision. Several adaptive techniques are used to improve the flexibility and robustness of the conventional spectral matching algorithm. An objective evaluation of the new voicing algorithm is obtained by comparing to manually estimated voicing values. Subjective tests of a sinusoidal coder using the new voicing algorithm show significantly better performance than the standard spectral matching under both clean and noisy environment.

I. INTRODUCTION

Voicing information plays an essential role in many low-bit rate speech coders. It is well known that the pure periodic excitation is a main cause of distortion in the traditional LPC vocoder. A mixed excitation model was introduced in the newer LPC type vocoders including the new MELP standard [1] in order to alleviate this problem. In the Multi-Band Excitation (MBE) model [3], a voiced/unvoiced (V/UV) decision is assigned to a group of harmonics of the fundamental frequency, and several V/UV decisions are transmitted to the decoder to generate synthesized speech which contains both harmonic and noise components. A number of low-bit rate coders including the sinusoidal transform coding (STC) coder [4] use only one parameter called transition frequency, or cutoff frequency, ω_c, instead of multi-band decisions to represent the voicing information. In such a model, the frequency band below ω_c is declared as voiced, and the frequency band above ω_c is declared as unvoiced. This simplified model gives similar performance with fewer bits compare to the multi-band model. In all these algorithms, the mixed excitation model and the voicing estimation technique are important factors which affect the subjective quality of the speech reproduction, particularly in acoustic background noise environment

Time-domain and frequency-domain techniques have been used to estimate the cutoff voicing frequency, ω_c (called below also simply "voicing"). In the time-domain approach, the speech signal is filtered into several frequency bands and a voicing decision is made in each band according to the maximum autocorrelation value around the pitch lag for that particular band [1, 2]. The time-domain method produces good results with low complexity if the number of bands is not large. In the frequency-domain technique [3, 4], the speech spectrum (or the residual signal spectrum) is matched to a harmonic synthetic spectrum model. If the matching is "good enough" for a frequency band (according to an empirical criterion), that frequency band is declared as harmonic-like (voiced), otherwise the band is declared as noise-like (unvoiced). An alternative frequency-domain method is based on declaring as voiced the bands characterized by equally spaced spectral peaks [5].

In this paper, we propose a new algorithm for estimating the voicing frequency, ω_c, using waveform time warping combined with spectral matching. Time warping is used to reduce the effect of pitch variations on the voicing estimation. The algorithm computes voicing decision for each harmonic in frequency domain. Then, a transition frequency is determined from the voicing information of all harmonics. The algorithm is flexible and can be used for producing either multi-band voicing decisions or a single voicing cutoff frequency.

In conventional spectral matching, the windowed speech spectrum is matched to the spectrum of the window function for each harmonic. This technique may

[*] This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Advanced Computer Comm., Cisco Systems, DSP Group, DSP Software Eng., Fujitsu, General Electric Company, Hughes Electronics, Intel, Nokia Mobile Phones, Qualcomm, Rockwell International, and Texas Instruments.

0-7803-4428-6/98 $10.00 © 1998 IEEE

fail because the speech spectrum around the harmonic peaks may not have the perfect shape of the window spectrum as a result of additive background noise and pitch variations. Although it is not very difficult for a human observer to judge whether a frequency range is dominated by harmonic components or noise just by observing the speech spectrum, a spectrum without ideal harmonic structure may cause problems for conventional spectral matching methods. Several techniques are introduced in the proposed algorithm to improve the flexibility and robustness of spectral matching including: window zoom, window position jitter etc. - see Section II.

The performance of the algorithm was evaluated using both subjective and objective criteria. For objective evaluation, the voicing information was obtained manually from the spectrum examination using visual tools and employed as a reference for determining a voicing error rate. For subjective evaluation, the algorithm was embedded into a harmonic coder [8] and compared to the standard spectral matching technique under both clean speech environment and noisy conditions.

II. Voicing Algorithm Description

1. Time warping for voicing estimation

A well known problem in voicing estimation is the effect of pitch period variation on the high frequency part of the speech spectrum: the harmonic structure may be vague in the high frequency range not because of low voicing but due to pitch variations. For rapid pitch changes, even if the speech has perfect periodicity, the spectrum at high frequencies may look similar to that of noisy (unvoiced) speech. This situation may result in errors in voicing estimation. We propose to solve this problem by using time warping with the objective to minimize pitch variations within the analysis frame.

Time warping has been previously used in speech coding with the objective of improving the efficiency of the adaptive codebook in a variation of the CELP coder called RCELP [6, 7]. In RCELP, the pitch-period contour of the original residual signal is modified by time warping to match a synthetic piecewise linear contour and this results in fewer bits being required to transmit the pitch delay. A piecewise linear pitch-period contour would not improve voicing estimation. To recover the harmonic structure in high frequency band, piecewise constant pitch-period contour is needed. The time-warping algorithm used to achieve a piecewise constant pitch-period contour is described below.

Assume the pitch information $P(t)$ of the original speech signal $x(t)$ is known. The objective is to transform the time-domain waveform $x(t)$, $t = t_i, \ldots, t_j$ into a signal $\xi(\tau)$, $\tau = \tau_i, \ldots, \tau_j$ which has constant pitch period P_m in the analysis window. Define the time-warping function $\zeta(t)$ by:

$$\zeta(t) = \frac{d\tau}{dt} \qquad (2)$$

The relationship between $x(t)$ and $\xi(\tau)$ is given by:

$$\xi(\tau) = \xi\left(\tau_i + \int_{t_i}^{t} \zeta(t)dt\right) = x(t) \qquad (3)$$

At time t, the pitch $P(t)$ should be modified to be P_m, hence the time-warping function is:

$$\zeta(t) = \frac{P_m}{P(t)} \qquad (4)$$

The modified signal $\xi(\tau)$ is output at times τ_i, τ_{i+1}, ... We assume the time-warping function is constant during the interval $[\tau_i, \tau_{i+1}]$ and given by

$$\zeta(t) = \frac{P_m}{P(t_i)} \qquad t_i \le t < t_{i+1} \qquad (5)$$

Introducing (5) into (3) we obtain the warping equation

$$\xi(\tau_{i+1}) = x\left(t_i + \frac{P(t_i)}{P_m}(\tau_{i+1} - \tau_i)\right) \qquad (6)$$

The continuous time pitch information $P(t)$ in (6) is obtained by assuming linear pitch variation between pitch estimation (update) points. The pitch period of the original speech signal is estimated every 80 samples. The analysis window (N = 240 samples) is centered at the middle of the current frame. Denote the pitch period of the current frame, i, as P_i. The constant frame pitch P_m is chosen to be the pitch period of current frame, eg. $P_m = P_i$. The pitch period for each sample in the analysis window is approximated as

$$P(n) = \begin{cases} P_{i-1} + \dfrac{(P_i - P_{i-1}) \cdot n}{N/2} & 0 \le n < N/2 \\ P_i + \dfrac{(P_{i+1} - P_i) \cdot (n - N/2)}{N/2} & N/2 \le n < N \end{cases} \qquad (7)$$

The above warping procedure is done frame by frame. The effect of the discontinuities at the frame boundaries is alleviated by using a Hamming window which is applied to the time warped signal before frequency analysis.

To illustrate the advantages of time warping, the original speech spectrum and the time warped modified speech spectrum for a typical voiced frame are shown in Figure 1. The modified spectrum has clearly better harmonic structure than the original spectrum.

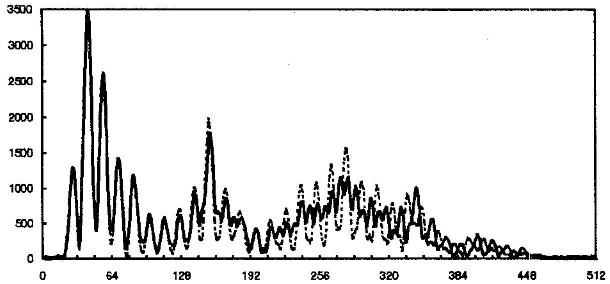

Figure 1. The original speech spectrum (solid line) and the speech spectrum after time-warping (dash line).

2. Spectral matching

This section reviews briefly a standard spectral matching technique. The speech spectrum is obtained for each frame using a 240 points Hamming window centered at the middle of the frame. The windowed signal is zero-padded to generate a 1024 point DFT $S(k)$, k = 0, 1, ..., 1023.

The speech spectrum in the neighborhood of each harmonic is matched to a harmonic spectrum (which is represented by the main lobe of the window spectrum) and to a flat spectrum. A harmonic of the speech spectrum is declared as voiced if the matching error for harmonic spectrum is smaller than that for the flat spectrum. Denote the fundamental frequency by f_0. The matching range for the lth harmonic is $[l{\cdot}f_0 - 0.5{\cdot}f_0,\ l{\cdot}f_0 + 0.5{\cdot}f_0]$. The corresponding range of the DFT coefficients is denoted by $[a_l, b_l]$. The matching error for the l-th harmonic is:

$$E(l) = \sum_{k=a_l}^{b_l} \left[|S(k)| - W(k) \right]^2 \qquad l = 1, \ldots, L \qquad (1)$$

where L is the largest harmonic number in the 4kHz frequency range. The matching spectrum $W(k)$ is either the main lobe of the window spectrum or a constant for each harmonic. If the length of $[a_l, b_l]$ is larger than that of the main lobe of window spectrum, the matching range is reduced to the length of the main lobe.

3. Adaptive spectral matching

The spectral matching technique presented in the previous section lacks robustness, particularly in background noise conditions. In order to improve the flexibility and robustness of the algorithm, the window main-lobe spectrum is modified for better matching as follows.

- The speech spectrum may not have very deep valley between harmonic peaks. This may cause large matching errors near the boundary of the window spectrum main-lobe, as the valleys of the window spectrum are deeper than those of the speech signal spectrum. To avoid this problem, the lowest level of the speech spectrum is estimated by interpolation between speech spectrum valleys and the window spectrum is hard limited to be no lower than the speech spectrum lowest level.
- The actual speech spectrum may have a wider main-lobe than the window spectrum. An adaptive zoom factor is introduced to widen the window spectrum to fit the actual speech spectrum. The maximum zoom factor is 1.5.
- The harmonic peaks of the actual speech spectrum may not appear exactly at the multiples of f_0. A jitter parameter is used to allow the peak of the lth-harmonic to move in the range $[l{\cdot}f_0 - 0.1{\cdot}f_0\ ,\ l{\cdot}f_0 + 0.1{\cdot}f_0]$.

These adaptive modifications increase significantly the flexibility and the robustness of the spectral matching algorithm.

An example which illustrates the advantages of adaptive spectral matching is shown in Fig. 2 and Fig. 3. The speech spectrum and the model window spectrum used in the spectral matching technique described in Section 2 are shown in Figure 2. It is easy to see that there is a significant mismatch between the two spectra due to the different depth of spectral valleys and some misalignment. Figure 3 presents the model window spectrum obtained using the adaptive spectral matching technique described in this section compared to the same segment of speech spectrum as that of Fig. 2. It is quite obvious that the spectrum obtained using the adaptive technique matches the speech spectrum much better.

4. Voicing Decisions

The U/V decision for each harmonic is based on comparing the ratio of harmonic matching error and flat-spectrum matching error with an adaptive threshold. For each harmonic if the error ratio is less than the threshold, the harmonic is declared as voiced, otherwise it is declared as unvoiced. The threshold is initially set to 1.0 and it is modified to be smaller if the high-band spectrum energy is much larger than the low-band spectrum energy, which means the speech is more likely to be unvoiced. Above the frequency of 2000Hz, the threshold is decreased gradually to favor an unvoiced decision.

Furthermore a single transition frequency can be determined from U/V decisions for all the harmonics. The transition frequency is set as high as possible provided the ratio of the number of voiced harmonics and the number of unvoiced harmonics below the transition frequency is higher than a frequency-dependent threshold.

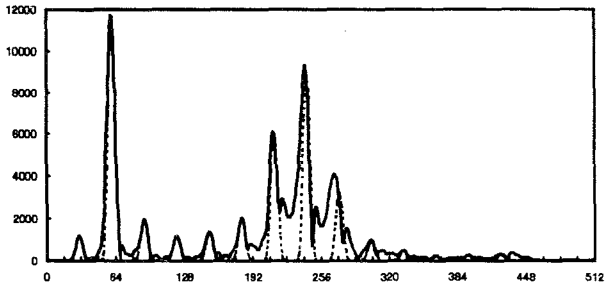

Figure 2. The original spectrum (solid line) and the harmonic model spectrum (dash line).

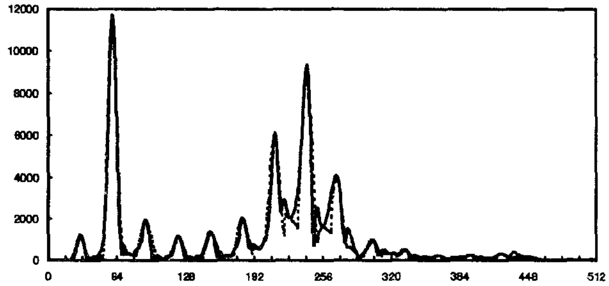

Figure 3. The original spectrum (solid line) and the modified harmonic model spectrum (dash line).

IV. Results

An evaluation of the objective performance of the proposed algorithm was done by comparing the cutoff frequency obtained from voicing program to the manually determined voicing number. For manual voicing determination, the spectrum of the speech signal was observed using a visual tool and a transition frequency was estimated each 10ms. Since the time-warping technique is not applied to the speech signal used for manual voicing determination, the voicing result obtained from time-warped speech is not compared to the manual voicing information.

The objective results, presented as a distribution of the normalized voicing error, are shown in Table 1. For both male and female speakers, the new algorithm reduces significantly the number of large voicing errors. The proposed voicing algorithm was embedded into a harmonic coder [8] for subjective quality testing. A conventional spectral matching based voicing algorithm similar to that used in MBE was also implemented for comparison. The new voicing algorithm was found to give better synthesized speech compared to the conventional method under both clean and noisy environment. The subjective tests indicate a preference of about 42.2% for the new voicing estimation versus 10.9% for conventional method with 46.9% of the same quality.

Table 1: Objective results for voicing algorithm

Conditions	Normalized voicing error	Conventional Method (%)	Adaptive Spectral Matching
Female	< 30%	80.77	86.41
	< 60%	10.57	9.89
	≥60%	8.66	3.70
Male	< 30%	67.85	79.42
	< 60%	15.62	14.05
	≥60%	16.53	6.53
Total	< 30%	74.17	82.84
	< 60%	13.15	12.01
	≥60%	12.68	5.15

References

[1]. A. V. McCree and T. P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," IEEE Trans. Speech and Audio Processing, vol. 3, pp. 242-250, July 1995.

[2]. R. L. Zinser, M. L. Grabb, S. R. Koch, and G. W. Brooksby, "Time Domain Voicing Cutoff (TDVC): A High Quality, Low Complexity 1.3-2.0 Kb/sec Vocoder," *Proceedings of the IEEE Speech Coding Workshop*, pp.25-26, Sept. 1997.

[3]. D.W. Griffin, J.S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36. pp.1223-1235, August 1988.

[4]. R. J. McAulay and T. F. Quatieri, "Sinusoidal Coding," Chapter 4 in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, eds., Amsterdam: Elsevier Science Publishers, 1995.

[5]. C. Papanastasiou, and C. S. Xydeas, "Efficient Mixed Excitation Models in LPC Based Prototype Interpolation Speech Coders," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.1555-1558, April 1997.

[6]. W. B. Kleijn, P. Kroon, and D. Nahumi, "The RCELP Speech Coding Algorithm," *European Transaction on Telecommunication*, vol. 5, pp. 573-582, Sept. 1994.

[7]. W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "Generalized Analysis-by-synthesis Coding and Its Application to Pitch Prediction," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I337-I340, 1992.

[8]. E. Shlomot, V. Cuperman, and A. Gersho, "Hybrid Coding of Speech at 4 kbps," *Proceedings of the IEEE Speech Coding Workshop*, pp.37-38, 1997.