MULTIPLE SOURCE MOS EVALUATION OF A FLEXIBLE LOW-RATE VOCODER

Richard L. Zinser, Mark L. Grabb, Steven R. Koch

GE Corporate Research & Development
One Research Circle
Niskayuna, NY 12309, USA

ABSTRACT

This paper describes the design and MOS performance of a family of low rate, low complexity speech coding algorithms known as Time Domain Voicing Cutoff (TDVC). TDVC is a predictive coding algorithm that employs a single transition frequency dividing voiced and unvoiced excitation. It provides the voicing flexibility of a frequency domain algorithm with lower complexity and rate overhead. A number of algorithm variants were MOS tested using three distinct sets of source material. The results are discussed in terms of performance for each of the three sources, and demonstrate that choice of source material has a great impact on both vocoder scoring and ranking.

1. INTRODUCTION

Very low rate speech coding (sub 2.0 kb/s) is an area where increasing levels of research effort are directed. The advent of handheld satellite telephony terminals has produced a need for robust speech and channel coding algorithms, while the demands for efficient system utilization provide motivation for a combined speech and channel rate below 4 kb/s. These developments in the commercial marketplace indicate a strong need for high-quality speech coding algorithms with a source rate of 1.3 to 2.0 kb/s.

At these low rates, a waveform-preserving coder (such as CELP) cannot perform well due to the high subframe rate required for reproduction of voiced speech. For this reason, our development work has concentrated on non-waveform preserving vocoders, with the goals of efficient quantization and low complexity. To best achieve these goals, we searched for a simple time domain speech production model that would provide the voicing flexibility of a frequency domain model (e.g. Multi-Band Excitation [1]), but with less complexity and fewer bits needed to transmit the voicing information. Working in the time domain also allows us to take advantage of efficient predictive spectral quantization techniques (e.g. Line Spectral Frequency Vector Quantization, or LSF-VQ [4]). The resulting speech coder algorithm is called Time Domain Voicing Cutoff (TDVC).

Because the TDVC algorithm is very flexible, a large number of algorithmic variations is possible. These variations allow TDVC to be tailored to a large number of environments. In the remainder of this paper, we describe a "baseline" version of the algorithm and several variations over rate, quantization, and enhancement algorithms. All variants were subjected to extensive MOS testing and analysis.

This work was funded by Lockheed Martin Corporation.

2. TDVC BASELINE

2.1 Concept

The baseline TDVC speech synthesizer consists of an excitation generator connected to an all-pole synthesis filter and an adaptive postfilter. Although the basic building blocks are similar to an LPC-10 or CELP synthesizer, the excitation generator is significantly different.

The underlying concept of TDVC is that there exists a single transition frequency (the voicing cutoff frequency) below which voiced excitation (e.g. periodic bandlimited pulses or a sum-of-sinusoids) is employed, and above which unvoiced excitation (bandlimited Gaussian noise) is employed. Quantizing this frequency is very simple. We have found that good performance can be attained with 8 equally spaced voicing cutoff frequencies. Thus, a total of 3 bits per frame are required for transmission. With the voicing information updated every 20 msec, a voicing transmission rate of 150 b/sec is generated.

The TDVC algorithm has more efficient voicing transmission than that used in MBE-type coders; the MBE coders make separate voicing decisions for several bands, and can use up to 11 bits (per frame) for quantization. It is also more efficient than a MELP [3] style coder, which requires 4 (bandpass voicing) + 1 (overall voicing) = 5 bits per frame (for a 4 band system operating at 2.4 kb/s).

The voicing cutoff frequency can be determined via a number of different techniques. A filter bank approach using pitch lag autocorrelation analysis similar to that of MELP may be employed, or a frequency domain (FFT) analysis similar to that of MBE could be used. The key differences are: 1) unlike MELP, there is no mixing of voiced and unvoiced excitation in the same frequency region, and 2) unlike MBE, there are not multiple “bands” of voiced and unvoiced excitation.

We have found that the filter bank approach can produce good results with low complexity. Nominally, a filter bank approach produces normalized autocorrelations calculated near the pitch lag for each band. In a MELP coder, these autocorrelations are used to calculate the bandpass voicing strengths. In a TDVC coder, the autocorrelations are used to determine the voicing cutoff frequency. To determine the cutoff frequency, a search is

performed over the autocorrelation array, and any band having a correlation greater than 0.6 is marked as voiced. The voicing array is then smoothed working from the lowest frequency band upward in the following fashion: an unvoiced band is marked voiced if it lies between two voiced bands; after 2 contiguous unvoiced bands are encountered, the remaining bands are marked unvoiced. In addition, band 0 may be marked voiced if band 1 is voiced. The cutoff frequency is selected at the frequency boundary between the last voiced and first unvoiced bands. If all bands are unvoiced, the cutoff frequency is 0 Hz; if all are voiced, the cutoff frequency is one-half the sampling frequency.

Some additional smoothing of the voicing cutoff frequency is performed to handle occasional irregularities in the periodicity of voiced speech. This smoothing takes into account past and future values of the input RMS power level, the zero crossing rate, the prediction gain, and the overall autocorrelation at the pitch lag. There are also special cases for determining the voicing cutoff frequency during plosive onsets.

2.2 Spectral Coding

A tenth-order LPC model is used to capture the short-time speech spectrum. The autoregressive coefficients are converted to LSFs prior to quantization. LSFs are computed every 20 milliseconds using Hamming weighted speech frames. In the synthesizer, the LSFs are interpolated every 5 msec before conversion to autoregressive coefficients to form the all-pole synthesis filter.

We have experimented with many different techniques for encoding the LSFs, which all fall into two broad classes. For TDVC coders operating in the 1.75-2.0 kb/s range, we vector quantize each frame's LSFs independently. For TDVC coders operating in the 1.3-1.75 kb/s range, we jointly encode the LSFs for two adjacent frames; typically, one of the frames will be vector quantized and one will be interpolated.

We perform time domain smoothing of the LSFs of adjacent frames in order to improve the evolution of the power-spectrum envelope over time. The Voronoi region-based approach described in [2] is our preferred approach.

2.3 Excitation

The all-pole filter is excited using a time domain waveform consisting of the sum of the voiced and unvoiced excitations. The voiced excitation is formed by a series of lowpass periodic pulses, while the unvoiced excitation consists of highpass Gaussian noise. The bandwidth of the pulse is set to the voicing cutoff frequency, while the noise is high-pass filtered in a complementary fashion.

A single gain is transmitted representing the rms value of the combined excitation. This gain is encoded with a 5-bit Lloyd-Max scalar quantizer.

The excitation signal is constructed on a pitch epoch-by-epoch basis. For each new epoch, a single bandlimited pulse is generated. During an epoch, all the parameters of the excitation are held constant: the pitch period (length of the epoch), the fundamental frequency of the voiced excitation, and the voicing cutoff frequency. The parameter values are determined at the beginning of the epoch by linearly interpolating current and previous frames' values according to the time position in the synthesis frame. Although this interpolation introduces a half-frame delay in the synthesized speech, it is critical for producing high quality output.

The pulsed excitation can be generated in a number of ways. One method would be to use the impulse response of a low-pass filter whose 3 dB cut-off frequency is equal to the voicing cutoff frequency. Another method would be to build up a bandlimited pulse by summing sinusoids that are harmonics of the fundamental frequency from the fundamental frequency up to the voicing cutoff frequency. In either case, the pulses must be normalized so that a unit variance results when the bandlimited noise is added. A suitably normalized pulse can be expressed by

$$epoch(i) = \sqrt{\frac{2}{n}} \sum_{k=1}^{n} \alpha(k) \cos(k\omega_0 i + phase(k))$$

where *epoch(i)* is the *i*-th sample of the pulse, *n* is the number of harmonics (determined by fundamental frequency and voicing cutoff frequency), $\alpha(k)$ are the sinusoidal amplitudes (nominally set equal to 1.0), ω_0 is the fundamental frequency, and *phase(k)* is the (fixed) phase offset for the *k*-th harmonic. Enhanced performance may be obtained by attenuating the harmonic amplitudes in spectral valleys.

2.4 Pitch Analysis

Any multi-frame smoothed pitch tracking algorithm can be used for TDVC. Multiple frames are required to smooth out occasional pitch doublings. In addition, it is desirable for the tracker to return a fixed value (last valid pitch, or any fixed value that is unrelated to the lag associated with peak autocorrelation) during unvoiced speech. This minimizes false-positive voicing decisions in the voicing cutoff smoothing algorithm.

A 1000 Hz low pass filter is used to preprocess the input speech before pitch analysis. The pitch is coded with a 6-bit logarithmically spaced table with lags between 20 and 118 samples. The table is similar to that used in FS-1015 (Federal Standard LPC-10 vocoder).

3. TDVC VARIANTS

3.1 Gain Calculation

TDVC system gain has been calculated and applied via 3 methods: LPC residual RMS, post-synthesis RMS matching, and analysis-by-synthesis matching. The LPC residual RMS technique consists of measuring the RMS value of a single 20 msec frame of the input speech filtered by the all-zero LPC "analysis" filter. This gain is applied to the unity-variance excitation signal in the synthesizer, before the LPC synthesis-

and post-filters. The post-synthesis RMS matching algorithm quantizes the input signal's RMS at the analyzer, and applies a gain scaling to the output of the synthesizer's LPC filter to match the input level. The third technique (analysis-by-synthesis) generates a "sample" excitation sequence, applies all enhancements and synthesis filtering to generate a "sample" output sequence, and divides the input RMS by the "sample" RMS to arrive at a system gain. The analysis-by-synthesis gain is applied before the LPC synthesis filter in the synthesizer. The MOS results for residual and analysis-by-synthesis gains are compared in Section 4.

3.2 Post-Analysis-by-Synthesis Harmonic Amplitude Correction

Using the gain generated by the analysis-by-synthesis algorithm, a set of LPC harmonic amplitudes was formed by evaluating the synthesis filter transfer function at the harmonic frequencies ω_k:

$$HA(\omega_k) = \left| \frac{G}{1 - \sum_i a(i) e^{j\omega_k i}} \right|$$

To form correction ratios, a 512 point, zero-padded FFT was taken of 256 input signal samples centered around the LPC analysis frame. A Hamming window was employed. The resulting FFT was peak searched near anticipated harmonic frequencies. FFT "bins" associated with each harmonic were identified, and harmonic amplitudes were calculated from each set of bins. The amplitudes were adjusted for the effect of the Hamming window, and correction ratios were formed by dividing the FFT amplitudes by the LPC amplitudes. Note that while these ratios correct errors in the LPC analysis, they also correct gain errors due to the analysis-by-synthesis framework.

In an effort to evaluate the effect of a quantizable set of corrections, the first 10 of these amplitude corrections were applied to the first 10 harmonics in the synthesizer (less than 10 were used if the voicing cutoff frequency so dictated). Linear interpolation was used to update the corrections for each pitch epoch. Results for the unquantized correction ratios are given in Section 4.

3.3 LSF Spectral VQ

The baseline for TDVC with independent LSF quantization is a 26-bit 3-split VQ, while the baseline for variants using LSF interpolation is a 30-bit 3-split VQ. Other methods used in our experiments include a 25-bit 4-stage VQ, and a 21-bit predictive + safety-net VQ [5].

3.4 Low Rate Spectral Coding Interpolation Approaches

Two approaches were used to code the LSF information for operation at 1500 b/sec and below. The first approach, alternate frame interpolation, consists of encoding odd frames with full quantization (split or multistage), and encoding even frames with a selectable weighted sum of future and past odd frames. Two bits are used to represent four interpolation weights.

The second approach, known as switched frame interpolation, consists of encoding two 20 msec LSF frames at a time. It is similar to the first approach, except an extra bit is sent once every 40 msec to indicate which frame is fully quantized and which frame is interpolated. If the first frame is designated interpolated, its future and past frames are used to form the weighted sum; if the second frame is designated interpolated, the two closest past frames are used to form the sum. The algorithm for selection of the coded and interpolated frames is still under development; presently a log sum of frame RMS and inverse spectral distortion is used. MOS results for both approaches are given in Section 4.

3.5 Bass Enhancement

Bass enhancement was introduced by increasing the power of the first three harmonics during voiced excitation. The increase was a function of pitch, with larger enhancements applied to lower-pitched speakers. The enhancement did not exceed 3dB.

4. MOS TESTING

Multiple MOS tests were conducted on TDVC variants and another commercially-available speech coding algorithm. These tests were structured to assess the effects of: 1) variations in test score due to different MOS sources; 2) variations in base coder rate (1375-2000 b/sec) due to the spectral LSFVQ algorithm; and 3) variations in spectral shaping, including formant and bass enhancement; and 4) variations in gain calculation.

For these tests, 3 sources suppliers were used: 1) ARCON, 2) COMSAT/BNR, and 3) DAM (Dynastat/DoD). All sources contained 3 male and 3 female talkers in a quiet background.

4.1 MOS Experiment 1 Results

The first test consisted of 7 speech coder variants, 6 MNRU conditions and the input source. Relevant coder conditions included 1) the base 2000 b/sec TDVC (26-bit split LSFVQ analysis-by-synthesis gain, and standard bass and formant enhancement), 2) 1950 b/s TDVC (same as base, but with 25-bit multistage LSFVQ), 3) 1375 b/sec TDVC (using alternate frame 3-bit LSF interpolation and 25-bit multistage LSFVQ on the non-interpolated frames), and 4) a commercially-available 2.4 kb/sec MBE coder. The MOS results are shown in table 4.1.

Table 4.1

#	Coder	MOS Source		
		ARCON	COMSAT	DAM
1	2000 TDVC baseline	3.17	3.04	2.89
2	1950 TDVC	3.25	2.97	2.98
3	2400 MBE	3.27	3.44	2.88
4	1375 TDVC	3.15	2.71	2.82

Several interesting conclusions can be drawn from the data in Table 4.1. The first is that the MOS score is not dramatically reduced when the rate is lowered below 2.0 kb/sec; in fact, the 1375 b/sec version does quite well using the ARCON source. The second, and more significant conclusion is that the source material has a direct and conclusive effect on the outcome of the test. At first look, it is clear that the ordering of the coders according to results is different depending on the source employed. A Newman-Keuls analysis of mean differences reveals more information. The COMSAT source is the only source that statistically separates the 2400 MBE coder from the 3 TDVC coders. Of particular interest is the comparison of the 2400 MBE and 1375 TDVC algorithms. Differences of 0.12, 0.73, and 0.08 are observed for the ARCON, COMSAT and DAM sources, respectively.

4.2 MOS Experiment 2 Results

The second MOS experiment was designed to assess the effect of variations in the TDVC algorithm. The test consisted of 13 speech coder variants, 6 MNRU conditions, and source. The ARCON and COMSAT sets of source material were the same as used in experiment 1; while the DAM source materials were processed with a different decimation filter.

Relevant speech coder conditions included: 1) base 2000 b/sec TDVC (26-bit split LSFVQ analysis-by-synthesis gain, and standard bass and formant enhancement), 2) coder #1 without any bass enhancement (BE), 3) 1500 b/sec TDVC (using alternate frame 2-bit LSF interpolation and 30-bit split LSFVQ on the non-interpolated frames, gain derived from the LPC residual, and no BE), 4) coder #3 with gain derived via analysis-by-synthesis, 5) 1525 b/sec version of coder #4 with switched interpolation and no BE, 6) coder #5 with the new version of BE, 7) coder #4 with unquantized post-gain harmonic amplitude corrections for (up to) the first 10 harmonics, and no BE, and 8) 1750 b/s TDVC with predictive LSFVQ, with BE. The results of MOS experiment 2 are shown in Table 4.2

Table 4.2 also provides some interesting conclusions. It appears that the standard bass enhancement algorithm provides some numeric improvement (~0.1 MOS) for ARCON and COMSAT sources, but the difference may not be statistically significant. On the other hand, the improved bass enhancement algorithm gives higher scores for all three sources (0.11 - 0.16), but still may not be statistically significant.

Analysis-by-synthesis gain does not seem to provide any advantage over gain derived from the LPC residual for the 1500 b/sec coders. In addition, it shows increased artifact content for very high-pitched speakers.

Switched frame interpolation (comparing coders 4 and 5) appears to offer no advantage over alternate frame interpolation.

It is interesting that coder #7 (1500 b/s with harmonic amplitude corrections) produced the highest TDVC score on the COMSAT source. This may be indicative that frequency domain amplitude fluctuations for adjacent harmonics (a phenomenon observed in the COMSAT source) are not modeled well by an autoregressive system.

Table 4.2

#	Coder	MOS Source		
		ARCON	COMSAT	DAM
1	2000 TDVC baseline	3.35	3.32	3.09
2	2000 baseline, no BE	3.24	3.22	3.10
3	1500 residual gain, no BE	3.15	3.06	3.06
4	1500 analysis-by-synthesis gain, no BE	3.16	3.14	2.94
5	1525 no BE	3.12	2.92	2.98
6	1525 improved BE	3.26	3.08	3.09
7	1500 harmonic amp. Correction, no BE	3.28	3.34	3.10
8	1750 predictive LSFVQ	3.35	3.17	3.08

Finally, the 1750 b/sec version of TDVC (predictive LSFVQ) appears to show little or no loss from the 2000 b/sec baseline for the ARCON and DAM sources, and a 0.15 MOS loss for the COMSAT source. For the majority of sources, this technique offers a significant reduction in rate with no apparent quality loss.

5. CONCLUSIONS

We have presented several algorithmic improvements to the TDVC family of coders, along with MOS test results measuring their effectiveness for various sources. Our results demonstrate that the choice of source material is a very important factor both in how coders score and in how they are ranked in testing. These results also suggest that MOS source material should be matched to the application for which the coder is intended.

6. REFERENCES

[1] Griffin D. and Lim J., "Multi-Band Excitation Coder," *IEEE Trans. on ASSP*, Vol 36, No. 8, pp. 1223-1235, 1988.

[2] Knagenhjelm H. and Kleijn W., "Spectral Dynamics is More Important than Spectral Distortion," *Proc. ICASSP*, pp. 732-735, 1995.

[3] McCree A., Truong K., George E., Barnwell T., Viswanathan V., "A 2.4 kb/sec MELP Coder Candidate for the new U.S. Federal Standard," *Proc. ICASSP*, pp. 200-203, 1996.

[4] Paliwal K. and Atal B., "Efficient Vector Quantization of LPC Parameters at 24 bits/frame," *IEEE Trans. on Speech and Audio Processing*, Vol. TSAP-1, pp. 3-14, 1993.

[5] Skoglund J. and Linden J., "Predictive VQ for noisy channel spectrum coding: AR or MA?," *Proc. ICASSP*, 1997.