# Extraction of LP-Based Features from One-Bit Quantized Speech Signals for Recognition Purposes

M. Felici, A. Ferrari, M. Borgatti and R. Guerrieri *
DEIS, Università di Bologna
Viale Risorgimento 2, 40136 - Bologna
ITALY

**ABSTRACT**

*A simplified fixed-point computation of cepstral coefficients, based on linear predictive analysis and infinite clipping of speech signals, is described. The autocorrelation function of the clipped signal is directly used to compute the linear predictor coefficients.*
*The performance of an isolated word recognition system based on these coefficients is presented and compared with a system which uses standard linear predictive cepstral features. The results show that these coefficients can be efficiently used for small dictionary speech recognition systems and, since the analog-to-digital conversion can be avoided, they are suitable for a low-voltage and low-power hardware implementation.*

## 1 INTRODUCTION

Single-word speech recognition systems are finding important applications in large volume products such as cellular phones and toys. In these fields, keeping the cost of the system low and reducing the total dissipated power is essential to meet the expectations of the end users. For these reasons, cheap VLSI recognition systems must keep low the computational requirements to reduce area and power consumption and should perform well when very simple analog/digital conversion circuits are used. In this work, we propose a very simple algorithm that exploits the fact that the intelligibility of speech signals is not severely affected by infinite clipping [1], i.e. one-bit quantization (OBQ).
Although in a "speech-to-text conversion" context the intelligibility of the signal under recognition is the basic requirement, little has been done to exploit a simple one-bit quantization of the speech signal to reduce the computational effort involved with the feature-extraction process.

In [2] it is proven that clipped speech preserves much of the discrimination power needed for recognition, showing that a speech recognition system based on optimal linear-predictor (LP) has a small degradation in recognition rate when the autocorrelation function (ACF) is not exactly computed from the original speech signal, but estimated from clipped signal.

In this paper we present a complete feature extractor for speech recognition systems which performs LP analysis on the clipped speech, then computes a set of variance-weighted cepstral coefficients for recognition. Under comparable conditions we reach the same overall performance in recognition rate as in [2], showing that clipped speech can be directly used in small vocabulary, simple speech recognition systems to save computations. Compared with the usual LP analysis, our OBQ-based LP analysis requires just one seventh of the global amount of multiplications and about a half of the number of additions. Compared with the ACF estimation method proposed in [2], our technique is significantly simpler since the repeated evaluation of a cosine function during each frame is not needed and the use of a frame distance measure simpler than Itakura-Saito is employed.

## 2 LP ANALYSIS

The short-term $p$-th order autocorrelation of the speech signal $s(i)$, evaluated on a $N$-samples wide window, can be computed as follows:

$$r_k = \frac{1}{N}\sum_{i=1}^{N} s(i)w(i)s(i+k)w(i+k) \qquad k = 0,\ldots,p \quad (1)$$

where $w(i)$ is a function that is zero out of window under examination. In case of clipped speech (i.e. $s(i) \in \{-1, 1\}$) and rectangular windowing function, the $k$-th autocorrelation coefficient can be simply computed by counting how many sign changes occur between samples belonging to the speech window at a distance equal to $k$. If $Z_k$ is that value, the autocorrelation function can be written as:

$$r_k = \frac{(N-k) - 2Z_k}{N} \qquad k = 0,\ldots,p \quad (2)$$

If we extend the counting of different samples over N comparisons (i.e. borrowing $k$ samples from the following window) the new counter value $\tilde{Z}_k$ is related to $Z_k$

*This work has been supported by SGS-THOMSON Microelectronics.

by the following:

$$Z_k \cong \frac{(N-k)}{N}\tilde{Z}_k \qquad k = 0, \ldots, p \tag{3}$$

where we assumed the stationarity of the clipped speech process over the window time duration. Extending the counting over N comparisons is useful when the analysis is performed on overlapping windows and the frame duration is an integer divider of the window duration. In this case a relevant reduction of computations can be achieved splitting the counting on single frames rather than counting over the whole window. The value of $\tilde{Z}_k$ is then straightforwardly obtained as sum of frame-based counts and these values can be reused in the evaluation of the next overlapping windows.
Using (3), the autocorrelation function can be estimated by the value of $r'_k$ as follows:

$$\begin{aligned} r'_k &= \frac{(N-k) - 2\tilde{Z}_k(1-\frac{k}{N})}{N} \\ &= \frac{(N-2\tilde{Z}_k)(1-\frac{k}{N})}{N} \qquad k = 0, \ldots, p \end{aligned} \tag{4}$$

Since the order of the autocorrelation is always much smaller than the window size ($p \ll N$), the second term in (4) can be neglected leading to the simple autocorrelation estimation formula:

$$r''_k = \frac{N - 2\tilde{Z}_k}{N} \qquad k = 0, \ldots, p \tag{5}$$

where the division is not essential and, when $N$ is a power of two, $Z_k$ directly converts in a fixed-point notation. Figure 1 shows the good accuracy of the autocorrelation estimators $r'_k$ and $r''_k$, based on $\tilde{Z}_k$, plotting the statistical distribution of the error measure:

$$e = \frac{\sum_{k=1}^{p}(r_k - r'_k)^2}{\sum_{k=1}^{p}(r_k)^2} \tag{6}$$

as observed over a large amount of recorded speech data. Equation (5) shows that the short-term autocorrelation of a one-bit quantized speech signal can be efficiently estimated by means of $p$ counters of sign changes over each frame period.

In [2] the short-term autocorrelation of the one-bit quantized speech signal is used to estimate the autocorrelation of the original speech signal using the Van Vleck's formula [3].
Unfortunately, this approach has several problems. First, Van Vleck's formula provides a poor estimation when, as in the case of speech signals, the hypothesis of stationarity and Gaussian statistics is far from the reality. Second, fundamental properties of the estimated autocorrelation function such as positive definiteness of the autocorrelation matrix are not guaranteed for this signal.

The Levinson-Durbin (LD) recursion is used to compute the LP coefficients from the ACF. This algorithm

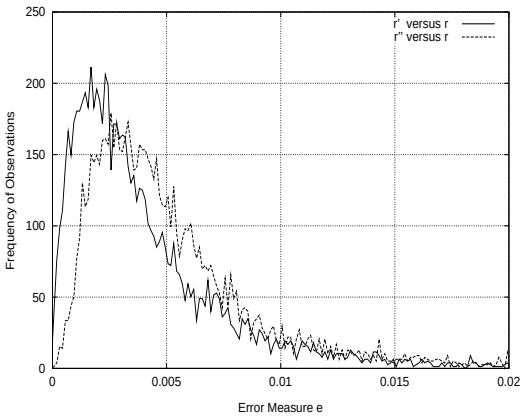


Figure 1: Observed distributions of the error measure $e$ over 10000 speech frames from the TI-46 database.

relies on the positive definiteness of the autocorrelation matrix which is a sufficient condition to guarantee the stability of the LP-inverse filter [4].
Therefore, ACF-estimators that do not preserve this property pose several numerical problems to this class of resolution methods.
One popular way to correct the ACF to make positive definite the autocorrelation matrix is to increase the 0-*th* order autocorrelation coefficient, $r_0$, multiplying it by a factor $(1+\lambda)$, $\lambda > 0$ as suggested in [5]. This stabilization method is physically equivalent to adding uncorrelated noise to the speech signal before performing the LP analysis.

Spectra in figure 2 show that much of the information of the speech signal is preserved after clipping. These plots report the filter spectra obtained from a LP analysis (12-*th* order filter) on Hamming windowed original speech and filter spectra obtained from a LP analysis (16-*th* order filter) on OBQ speech. Same window duration ($32ms$) is used in both cases. As stated in [6], if clipping is done after high-pass filtering speech, the signal obtained is more intelligible than simple clipped speech signal. For this reason preemphasis is used.

## 3 CEPSTRAL COEFFICIENT RECURSION

Cepstral coefficients can be computed directly from the LP model as shown in [5]. If we call $c_i$ the *i-th* cepstral coefficient and $a_i$ the *i-th* LP coefficient, a recursive formulation of the computation is:

$$\begin{aligned} c_1 &= -a_1 \\ c_i &= -a_i - \sum_{j=1}^{i-1}(1-\frac{j}{i})a_j c_{i-j} \end{aligned} \tag{7}$$

By reformulating the recursion with the position:

$$\xi_i = -i\ c_i \tag{8}$$

a substantial saving in the number of multiplications can be achieved. Simply by multiplying the second relation

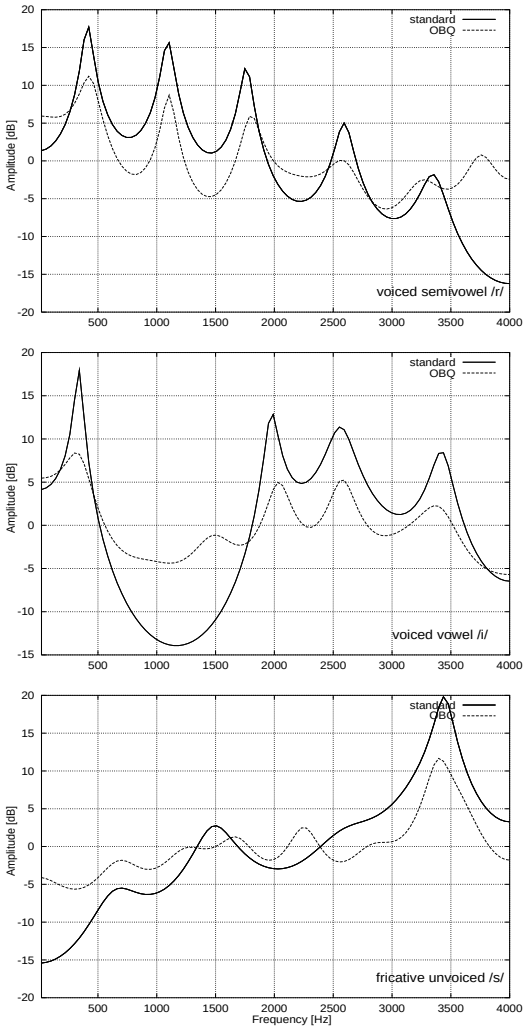


Figure 2: Comparison of LP inverse filter spectra obtained from a LP analysis on OBQ and original speech signals.

in (7) by $i$ and exploiting (8) we obtain:

$$\begin{aligned} \xi_1 &= a_1 \\ \xi_i &= i\ a_i + \sum_{j=1}^{i-1} a_j \xi_{i-j} \end{aligned} \tag{9}$$

For 15 cepstral coefficients this formulation reduces the number of multiplications to 57% of those in the straightforward implementation of (7).

Note that a set of standard deviation weighted cepstra can be directly computed normalizing the set of $\xi_i$ coefficients by their standard deviation.
Such features have the same statistical influence in the subsequent frame distance computation.

## 4 IMPLEMENTATION

A fixed-point implementation of Levinson-Durbin recursion is given in [7]. It is shown that the computation

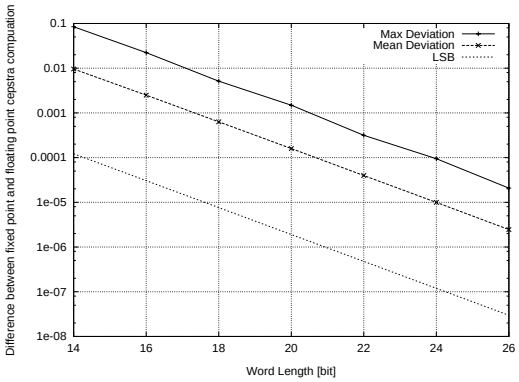


Figure 3: Statistics on $|\overline{c}_i - c_i|$ , in the case of the correct OBQ-ACF computation, are reported. $\overline{c}_i$ is the generic cepstrum calculated with fixed point n bit arithmetic while $c_i$ is its floating point version.

of inverse LP filter coefficients $a_n$ requires the use of a scaling factor $A$ to prevent overflows (following [7], A is fixed to 4).

The scaling of the cepstral recursion (7) is straightforward: since $|c_1|$ is equal to $|a_1|$ and cepstral coefficient ranges decrease with the coefficient index [8], the same scaling factor A can be used for the cepstral recursion.

The reformulated cepstral recursion (9) requires further scaling due to the fact that the quantities $\xi_i$ are greater than $c_i$. We found that, for a number of cepstral coefficients equal to 15, a further scaling by A is adequate. Therefore, calling $\overline{a}_i$ the *i-th* LP coefficient scaled by A and $\overline{\xi}_i$ the value of $\xi_i$ scaled by $A^2$, the fixed point implementation of (9) becomes:

$$\begin{aligned} \overline{\xi}_1 &= \overline{a}_1 / A \\ \overline{\xi}_i &= \frac{i}{A}\overline{a}_i + \sum_{j=1}^{i-1} \overline{a}_j \overline{\xi}_{i-j} \end{aligned} \tag{10}$$

The effect of word length on the accuracy of cepstral coefficient computation has been studied: 16 bit-wide integer arithmetic gives good results for recognition purposes. Good agreement with results presented in [7] has been found in terms of the occurrence of numerical instabilities versus word length.

The deviation of fixed point cepstral coefficients from floating point ones (in a logarithmic scale) versus word length is shown in figure 3. In the same figure is also shown the LSB value at the given word length assuming a fixed point range of $[-1, 1]$. The maximum numerical error with word length equal to 16 is about 0.02. However, experimental results show that only the six most significant bit of cepstral coefficients are needed for recognition purposes (LSB value in this case is about 0.03).

Table 1 reports computational requirements of standard and OBQ LP cepstrum algorithms tuned for best recognition rates of our system. Note that divisions,

required by Levinson-Durbin recursion, are not taken into account while multiplications required for cepstral coefficients weighting are counted in cepstral recursion multiplications and further multiplications required by windowing are counted for the standard system.

The computation of ACF does not require multiplications in the OBQ signal case. The use of a rectangular window and the fact that the frame duration is an integer divider of the window duration allow to reduce the number of required operations further if sign change computation is performed on a frame by frame basis rather than over each window. Only a counter running over the frame period per ACF coefficient is required. Therefore the number of sign changes over a window may be obtained as sum of past frame counting (counting that must be temporarily stored). Although LP-cepstral recursion requires a larger number of operations due to a higher inverse order in the OBQ case, the reduction of computations in ACF stage overcomes this drawback: the number of multiplications is reduced by a factor of 7.5, while the number of additions is about a half of the additions required by the standard system.

## 5 THE RECOGNITION SYSTEM

Tests have been based on "TI 46 Isolated Word Corpus". It contains 10+16 repetitions (divided in training and test set) of 20 words (the ten English digits plus ten general purpose commands) pronounced by 16 speakers (half male and half female). Signals have been filtered and down-sampled to $8kHz$. Words have been automatically end-pointed.
Two recognition systems based on LP-Cep features have been compared: one that relies on OBQ speech signal and a second that relies on original undistorted speech signal. Table 1 shows conditions under which best recognition rates for our recognition systems have been reached. In both cases signals are preemphasized before sampling.

We have found that Levinson-Durbin recursion shows instabilities when the estimation of the original signal ACF, according to Van Vleck's relation, is used as in [2]. To avoid this, the value of $\lambda$, that determines the enhancement of $r_0$, was set to 0.45.
The use of the approximation (5) also requires a stabilization, but, since estimation is quite good, a value of 0.1 for $\lambda$ has been used. With this value of $\lambda$ the problem of instability was solved. Moreover overflows have never been observed, proving that the scaling for the fixed point arithmetic implementation is correct.

Classification relies on a DTW algorithm based on $L2$ distance. Using a clusterization procedure that splits clusters until they contain only utterances of a single word, the results (in terms of recognition rates) obtained for the 20-words recognition system are described below.
**Speaker dependent test.** For each speaker, training set is clustered and then test set is classified with respect to the calculated centroids; recognition rate is 99.5% for standard cepstrum (1.3 centroids/word) and 98.8% for OBQ cepstrum (1.5 centroids/word).

| | OBQ-LP-cepstrum | | | | std-LP-cepstrum | | | |
|---|---|---|---|---|---|---|---|---|
| Oper. | ACF | LP | cep | Tot | ACF | LP | cep | Tot |
| ADD | 1088 | 257 | 105 | 1450 | 2405 | 145 | 55 | 2605 |
| MUL | 0 | 257 | 134 | 391 | 2610 | 145 | 76 | 2831 |

Table 1: Number of operation per frame for standard signal (window duration 24 ms, LP order 12, 11 cepstra) and OBQ signal (window duration 32 ms, LP order 16, 15 cepstra) analysis systems (frame duration of $8ms$ is assumed for both cases, Hamming windowing is used in the standard system).

**Multi-speaker test.** First four male and female speaker training sets have been merged, all these utterances have been clustered and then their test sets have been classified with respect to calculated centroids; recognition rate is 99.5% for standard cepstrum (6 centroids/word) and 98.9% for OBQ cepstrum (8 centroids/word).

The multi-speaker recognition rate becomes 97.2% when white noise at a SNR of 10dB is added to all the utterances in the database before training and classification processes. The clusterization procedure generates a larger number of centroids per word, about 14, in this case.

### References


[1] J.C.R.Licklider. Effects of amplitude distorsion upon the intelligibility of speech. *J. Acoust. Soc. Am.*, 18:429–434, 1946.

[2] V.Lipovac. Zero-crossing-based linear prediction for speech recognition. *Electronics Letters*, 25(2):90–92, 1989.

[3] D.Middleton J.H.Van Vleck. The spectrum of clipped noise. *Proc. IEEE*, 54:2–19, 1966.

[4] John Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(4):561–580, 1975.

[5] J.W.Picone. Signal modeling techniques in speech recognition. *Proc. IEEE*, 81(9):1215–1247, 1993.

[6] R.J. Niederjohn I.B. Thomas. The intelligibility of filtered-clipped speech in noise. *J. Audio Eng. Soc.*, 18(3):193–197, 1970.

[7] A.H.Gray J.D.Markel. Fixed point implementation of a linear prediction autocorrelation vocoder. *IEEE Trans. ASSP*, 22(4):273–282, 1974.

[8] J.Tohkura. A weighted cepstral distance measure for speech recognition. *IEEE Trans. ASSP*, 35(10):1414–1422, 1987.