Hybrid Scalar/Vector Quantization of Mel-Frequency Cepstral Coefficients for Low Bit-Rate Coding of Speech

Laura E. Boucheron, Phillip L. De Leon, and Steven Sandoval
Klipsch School of Electrical and Computer Engineering
New Mexico State University
Las Cruces, NM 88003 USA
{lboucher, pdeleon, xxspsxx}@nmsu.edu

ABSTRACT

In this paper, we propose a low bit-rate speech codec based on a hybrid scalar/vector quantization of the mel-frequency cepstral coefficients (MFCCs). We begin by showing that if a high-resolution mel-frequency cepstrum (MFC) is computed, good-quality speech reconstruction is possible from the MFCCs despite the lack of explicit phase information. By evaluating the contribution toward speech quality that individual MFCCs make and applying appropriate quantization, our results show perceptual evaluation of speech quality (PESQ) of the MFCC-based codec matches the state-of-the-art MELPe codec at 600 bps and exceeds the CELP codec at 2000–4000 bps coding rates. The main advantage of the proposed codec is in distributed speech recognition (DSR) since speech features based on MFCCs can be directly obtained from codewords thus eliminating additional decode and feature extract stages.

I. INTRODUCTION

The cepstral analysis of speech signals is a homomorphic signal processing technique to separate convolutional aspects of the speech production process [1]. Cepstral analysis allows the pitch and formant structure of speech to be easily elucidated which is important for pitch detection, phoneme recognition [2], [3], and speaker characterization [4], [5]. As such, cepstral analysis finds widespread use in speech processing including automatic speech recognition (ASR) and speaker recognition (SR). In particular, analysis based on the mel-frequency cepstrum (MFC) with a basis in human pitch perception [6], [7] is perhaps more common, e.g., [5].

Reconstruction of a speech waveform from mel-frequency cepstral coefficients (MFCCs) is a challenging problem due to losses imposed by discarding the phase spectrum and the mel-scale weighting functions. Among the earliest investigations for reconstruction of a speech waveform from MFCCs can be found in [8]. In this work, the authors propose an MFCC-based codec for use in distributed speech recognition (DSR) where MFCC feature vectors are extracted and quantized by the client before transmission over the network. This approach reduces system complexity since an alternate codec would require server-side decoding and extraction of MFCCs before ASR—with an MFCC-based codec, these latter two steps are unnecessary.

* Direct all correspondence regarding this manuscript to Dr. Phillip De Leon (pdeleon@nmsu.edu).

The challenge in the reconstruction of speech from an MFCC-based feature extraction process normally used in ASR (13-20 MFCCs per frame) is that too much information is discarded to allow a simple reconstruction of a speech signal [9]. One method we previously proposed in [10], reconstructs the speech waveform by *directly inverting* each of the steps involved in computing MFCCs. For the steps which impose losses, we use a least-squares (LS) inversion of the mel-scale weighting functions and an iterative LS phase estimation method. The key to this approach is to simply not discard too much information by using a high-resolution MFC (large number of MFCCs per speech frame), thus eliminating the need for auxiliary computation of fundamental frequency as needed in other methods [8], [9], [11], [12]. Prior to [10], this approach does not appear to have been proposed despite yielding a much simpler reconstruction algorithm than the sinusoidal-synthesis based methods presented in [8], [9], [11], [12].

Having developed a method to reconstruct good quality speech from a high-resolution MFC, we now show in this paper that through proper quantization of high-resolution MFCCs we can encode at 4800 bps rates (compatible with the ETSI Aurora DSR standard [11]) while at the same time enabling good quality, intelligible, reconstructed speech. This high-resolution MFCC vector can be easily downconverted to the standard low-resolution MFCC vector (13–20 coefficients per frame) for compatibility with ASR. We argue that our proposed approach satisfies the front-end DSR requirements: 1) ability to code MFCCs at standard bit-rates, 2) a simple downconversion to lower dimensional MFCC vectors compatible with ASR, and 3) good-quality reconstruction of the speech waveform from the MFCCs. We also show that the high-resolution MFC can be coded at bit-rates as low as 600 bps, yielding speech quality approaching that of the state-of-the-art MELPe codec [13]–[16]. At higher bit-rates, the MFCC-based codec yields speech quality better than that of CELP-based codecs [17].

This paper is organized as follows. In Section II, we review the procedure for reconstruction of the speech waveform from MFCCs previously described in [10]. In Section III we analyze and discuss the resulting perceptual artifacts due to the reconstruction. In Section IV, we describe the MFCC-based speech codec which utilizes the proposed reconstruction method and present results. Finally, we conclude in Section V.

II. Reconstruction of the Speech Waveform from Mel-Frequency Cepstral Coefficients

A. Cepstrum

Computation of the cepstrum begins with the discrete Fourier transform (DFT) of a windowed speech signal s:

$$x_r[m] \;=\; s[rR+m]w[m] \tag{1}$$

where w is the length L window ($0 \le m \le L-1$), R is the window or frame advance in samples, and r denotes the frame index. For convenience, we denote the speech frame as

$$\mathbf{x} \;=\; [x_r[0], x_r[1], \ldots, x_r[L-1]]^T \tag{2}$$

(we drop the subscript r to simplify notation) and the spectrum as the Discrete Fourier Transform (DFT) of $\mathbf{x}$

$$\mathbf{X} \;=\; \mathcal{F}\{\mathbf{x}\}. \tag{3}$$

The cepstrum of $\mathbf{x}$ may be defined as

$$\mathcal{C} \equiv \mathcal{F}^{-1}\left\{\log\left|\mathbf{X}\right|\right\} \tag{4}$$

where the inverse discrete Fourier transform $\mathcal{F}^{-1}$ is applied to the log-magnitude spectrum of $\mathbf{x}$.

B. Mel-Frequency Cepstrum

In the definition of Mel-Frequency Cepstral Coefficients (MFCCs) $\mathcal{M}$ we apply a set of weighting functions Φ to the power spectrum prior to the Discrete Cosine Transform (DCT) and log operations [7]

$$\mathcal{M} = \mathcal{DCT}\left\{\log\mathbf{\Phi}\left|\mathbf{X}\right|^2\right\}. \tag{5}$$

This weighting $\mathbf{\Phi}$ is based on human perception of pitch [6] and is most commonly implemented in the form of a bank of filters each with a triangular frequency response [7]. The mel-scale weighting functions ϕ_j, $0 \leq j \leq J-1$ are generally derived from J_1 triangular weighting functions (filters) linearly-spaced from 0–1 kHz, and J_2 triangular weighting functions logarithmically-spaced over the remaining bandwidth (1–4 kHz for a sampling rate of 8 kHz) [7], where $J_1 + J_2 = J$. Additionally, in our work we use two "half-triangle" weighting functions centered at 0 and 4 kHz which we include in J_1 and J_2 since these will directly affect the number of MFCCs. The use of the two "half-triangle" weighting functions improves the quality of the reconstructed speech waveform (described in the next section). In usual implementations, $J < L$ and thus this weighting may also be thought of as a perceptually-motivated dimensionality reduction.

The mel-weighted power spectrum in (5) can be expressed in matrix form as

$$\mathbf{Y} = \mathbf{\Phi}\left|\mathbf{X}\right|^2 \tag{6}$$

where $\mathbf{Y}$ is $J \times 1$, the weighting matrix $\mathbf{\Phi}$ is $J \times L$ and has columns ϕ_j, and $\left|\mathbf{X}\right|^2$ is $L \times 1$.

C. Reconstruction from MFCCs

The MFCCs are primarily used as features in speech processing and are not normally converted back to speech, however, an estimate of the speech frame can be made from the MFCCs. In (5), two sources of information loss occur: 1) application of the mel-scale weighting functions and 2) the phase spectrum is discarded in computing the power spectrum. Otherwise, the DCT, log, and square operations are all invertible. Thus, estimation of the speech frame from the MFCCs requires a pseudo-inverse of $\mathbf{\Phi}$ and an estimate of the phase spectrum.

1) Least-Squares Inversion of the Mel-Scale Weighting Functions: Since $J < L$ we are presented with an under-determined problem. In order to solve this problem, we use the Moore-Penrose pseudo-inverse $\mathbf{\Phi}^\dagger = \left(\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\mathsf{T}$ and form a LS solution, i.e., the solution of minimal Euclidean norm, for $\left|\mathbf{X}\right|^2$ as

$$\hat{\left|\mathbf{X}\right|^2} = \mathbf{\Phi}^\dagger\mathbf{Y} = \mathbf{\Phi}^\dagger\mathbf{\Phi}\left|\mathbf{X}\right|^2 \approx \left|\mathbf{X}\right|^2. \tag{7}$$

2) Least-Squares Estimation of Speech Frame from Magnitude Spectrum: After pseudo-inversion of the mel-scale weighting functions, we are left with a magnitude spectrum from which we must estimate the speech frame. In order to compute the inverse transform, we must estimate the phase spectrum since this is discarded during computation of the MFCCs. Due to the under-constrained nature of the pseudoinverse $\mathbf{\Phi}^{\dagger}$, it is important to note that $\hat{\mathbf{X}}$ will not necessarily be a valid STFT in the sense that an STFT contains inherent structure in time and frequency due in large part to the overlap of the windowing process [1]. Furthermore, the phase information of the DFT will not be available for reconstruction of speech. Even with these limitations, we can reconstruct the speech waveform with the "closest" valid STFT $\tilde{\mathbf{X}}$, in terms of least squared-error (LSE), via the well-known Least-Squares Estimate, Inverse Short-Time Fourier Transform Magnitude (LSE-ISTFTM) algorithm [1]. The LSE-ISTFTM algorithm iteratively estimates the phase spectrum and couples this to the given magnitude spectrum resulting (after inverse transformation) in a time-domain estimate of the speech frame. The complete speech waveform is then reconstructed via an overlap-add procedure from the sequence of estimated speech frames.

III. Quality of Speech Reconstructed from MFCCs

Although the DCT, log, and square operations in (5) are all invertible, we utilize a pseudo-inverse of the mel-scale weighting functions and a phase estimate (LSE-ISTFTM) in order to complete the reconstruction of the speech frame; these two steps will impose quality losses. We measure the quality of the reconstructed speech signal using the perceptual evaluation of speech quality (PESQ) metric. Although originally developed for evaluation of speech quality in telecommunications applications where speech coding and distortions due to network conditions degrade quality, PESQ has been used for other evaluation of other speech processing algorithms. In a recent study on objective quality measures for speech enhancement, researchers have found that of the seven most widely used objective measures tested, the PESQ measure yielded the highest correlation ($\rho = 0.89$) to overall subjective quality and signal distortion [18]. In this work, PESQ results were averaged over a sample of 16 TIMIT speakers (8 female and 8 male) downsampled to a rate of $f_s = 8000$ Hz; each signal is ~ 24s in duration . The baseline PESQ score for the TIMIT reference signals is 4.5.

We begin by computing J MFCCs as in (5) using a 240 sample (30 ms) Hamming window with a 120 sample frame advance (50% window overlap). The number of MFCCs over 0–1 kHz, J_1 is selected as follows. We set $J_1 = 30$ for $J \geq 60$ or for $J < 60$, J_1 is selected for highest PESQ ($J_1 = [7, 15, 20, 30, 30]$ for $J = [10, 20, 30, 40, 50]$ respectively), The number of MFCCs over 1–4 kHz, $J_2 = J - J_1$. For a 30 ms window length, the DFT resolution is $33\frac{1}{3}$ Hz, providing exactly 30 frequency points over 0–1 kHz. Thus, for $J \geq 60$ there is no binning of the first 1 kHz; equivalently, the upper 30×30 block of $\mathbf{\Phi}$ is identity. From the MFCCs, we reconstruct the speech waveform using the method described in Section II-C. Fig. 1 shows the PESQ as a function of J (the number of MFCCs) for several different values of LSE-ISTFTM iterations.

We see that quality of the reconstructed speech signal from $J \geq 40$ MFCCs is fair (~ 3.25 PESQ MOS) when the number of LSE-ISTFTM iterations is at least 50. With $J = 70$ and 500 LSE-ISTFTM iterations, quality is fair/good (~ 3.6 PESQ MOS) and with fewer than 40 MFCCs, quality degrades rapidly. We also note that for a large

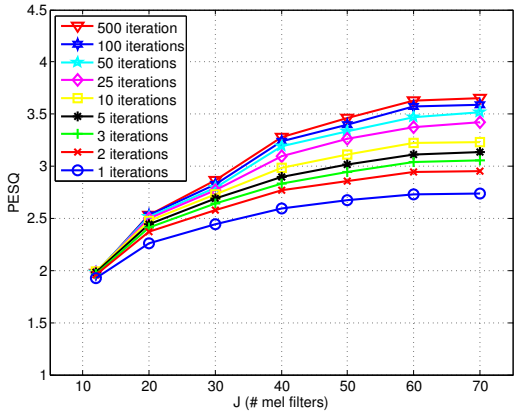


Fig. 1. Inversion of MFCCs in a clean phase-less environment at a various number of iterations. These results are averaged for a sample of 16 TIMIT speakers.

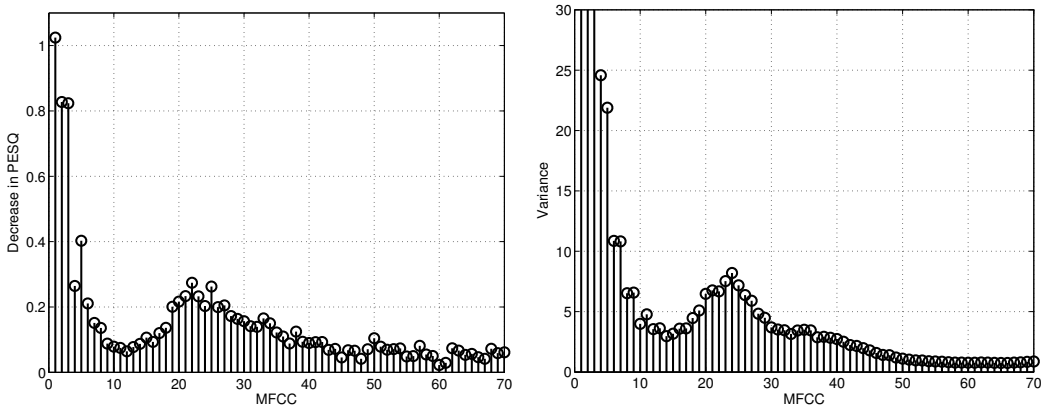
number of MFCCs ($J \geq 40$), doubling the number of LSE-ISTFTM iterations results in small PESQ improvement (~ 0.1 PESQ MOS point). Thus we find that the quality of the reconstructed speech from MFCCs depends more on resolution (number of MFCCs) than the number of LSE-ISTFTM iterations. For practical implementation with a large number of MFCCs, we find that 100 iterations provides a good balance between reconstruction quality and computation and yields a PESQ score within $\sim 2\%$ of the solution obtained with 500 iterations. For this reason, we will use a total of 70 MFCCs and the LSE-ISTFTM algorithm with 100 iterations for all work and when evaluating the MFCC-based codec. This yeilds a benchmark PESQ score of 3.58, thus the signal degradation imposed by MFCC computation in terms of PESQ is 0.92.

IV. Mel-Frequency Cepstrum-Based Speech Codec

In the previous sections, we have outlined a procedure to reconstruct speech frames from MFCCs and measured signal degradation imposed by MFCC computation. We now outline a method for quantization of the MFCCs for low bit-rate speech coding.

A. Assessing the Contribution of Individual MFCCs to Speech Quality

In order to determine a bit-allocation scheme for the MFCCs, we assess the relative contribution of individual MFCCs on speech quality. Tests were conducted in which we substituted (one at a time) a single MFCC with its mean value prior to reconstruction of the speech signal and computed PESQ. The mean values were computed using the entire TIMIT corpus and the tests were conducted using 16 TIMIT speakers. The resulting average *decrease* in PESQ is shown in Fig. 2(a). We see a large decrease in PESQ (significant degradation) when any of the first several (~ 7) MFCCs are replaced by their mean value; less significant degradation occurs for coefficients in the approximate range of 8–30, when a total of 70 MFCCs are used. This is not unexpected given the direct correspondence of the initial part of the MFC to formant structure (coefficients 1–14). The source of the smaller degradation is due to the appearance of pitch period (i.e., vocal excitation) information (coefficients 15–30); it is the high resolution MFC that allows for the appearance of pitch information as opposed to the standard lower



(a) Contribution to speech quality of individual MFCCs

(b) Variance of individual MFCCs

Fig. 2. Contribution to speech quality of individual MFCCs. In (a), the MFCC was replaced (one at a time) by the mean, speech signal reconstructed from MFCCs, and PESQ decrease was measured. In (b), the variance of individual MFCCs is measured.

resolution MFC. Overall, it appears that the most important MFCCs for perceptual quality of the reconstructed speech are the first several coefficients which correspond to formant structure. Fig. 2(a) suggests that more bits will have to be allocated to those MFCCs which contribute most to speech quality; other MFCCs may be discarded and substitution of the mean-value (stored in a lookup table) may be sufficient.

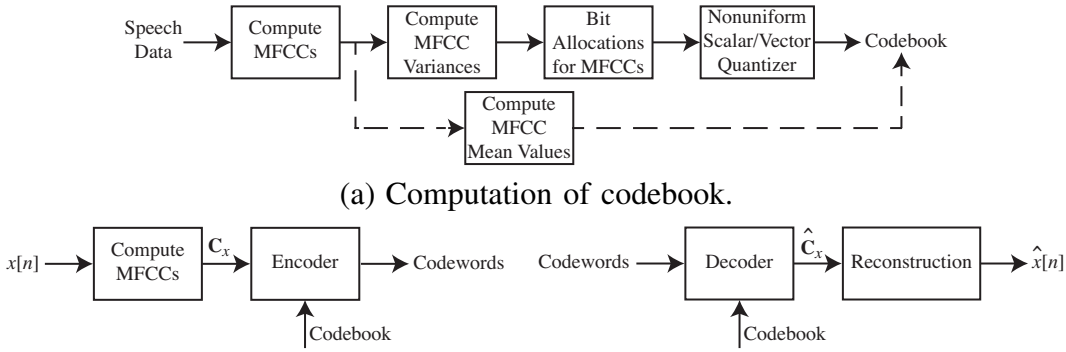
Shown in Fig. 2(b) is a plot of the variance of individual MFCCs across the speech frames. We see that coefficient variance is related to the individual coefficient importance. Computation of the variance of MFCCs is less computationally expensive and can therefore be computed over more speakers for a more accurate estimate of coefficient importance. It is this measure of coefficient importance that will be used in our proposed speech codec.

B. Non-Uniform, Scalar Quantization of MFCCs

We first consider non-uniform, scalar quantization (SQ) of the MFCCs. The non-uniform quantization levels are determined using the Lloyd algorithm (k-means clustering) [1]. Allocating a fixed 4 bits per MFCC, which yields a bit-rate of $4 \times 70 \div 0.015 = 18,667$ bps, we can realize a PESQ of 3.45 —only 0.13 PESQ MOS points below the reference which does not quantize the coefficients. This small degradation suggests 4 bits per MFCC are sufficient to code any MFCC with minimal loss.

In order to reduce the coding rate, we next consider reducing the number of bits per MFCC based on the variance as discussed in Section IV-A. Given a target bit-rate, we proportionally allocate bits to each MFCC according to the values shown in Fig. 2(b) allowing for a maximum of 4 bits and a minimum of 0 bits; in the latter case, we reconstitute the MFCC by using the coefficient's mean value (previously determined from speech data and stored in a lookup table at the decoder). Thus, the number of bits allocated to coefficient j is

$$B_j = B\sigma_j^2 / \sum_k \sigma_k^2, \tag{8}$$



(a) Computation of codebook.

(b) MFCC-based encoder.

(c) MFCC-based decoder.

Fig. 3. (a) Computation of codebook as outlined in Section IV where computation of the individual MFCC mean values is only used for scalar quantizer in Section IV-B, (b) MFCC-based encoder, and (c) MFCC-based decoder where the reconstruction block includes both the LS inversion of the mel-scale weighting functions and LSE-ISTFTM.

where B is the total number of bits per frame and σ_j^2 is the variance of the j-th MFCC. B_j is then rounded to an integer for implementation purposes.

Computation of the codebook is illustrated in Fig. 3(a), the blocks summarize the above information: from a set of speech data, we begin with computation of high-resolution MFCCs, measure mean and variance of individual MFCCs, determine the bit allocations according to (8) for the given bit-rate, and determine the scalar or vector quantization points, i.e., codewords. The proposed encoder is shown in Fig. 3(b), where the speech signal is windowed and MFCCs computed (Section II-B) and codewords are output. Finally, the decoder is shown in Fig. 3(c) where codewords are decoded to MFCCs according to the codebook and the speech frame is reconstructed (Section II-C).

The performance at various bit-rates for the proposed non-uniform, scalar-quantized MFCC-codec is shown in Fig. 4 (red circle solid line). The reconstructed speech is intelligible, and the most noticeable distortion is a muffling effect during voiced speech segments. This muffling effect is most likely caused by inaccuracies in the estimation of phase information which worsens at lower bit-rates. However, the reconstructed speech is free of the harsh synthetic sounds of many model-based codecs.

Interestingly, in the case of small overlap, we have found that inserting interpolated frames can improve quality of the decoded speech. These inserted frames are the direct linear interpolation of the two adjacent frames and are used by the LSE-ISTFTM algorithm as if they were a normally computed speech frame. Each interpolated frame essentially reduces the frame advance by a factor of 2. Fig. 4 illustrates the effect of inserting 3 interpolated frames for the SQ (red circle dashed line). For this case, recalling that the original signal was computed with 50% overlap, this is an approximation to a signal that was computed for 87.5% overlap. It is hypothesized that the redundancy of the interpolated frame improves in the inversion process of the LSE-ISTFTM algorithm which is a large source of quality loss (Section III).

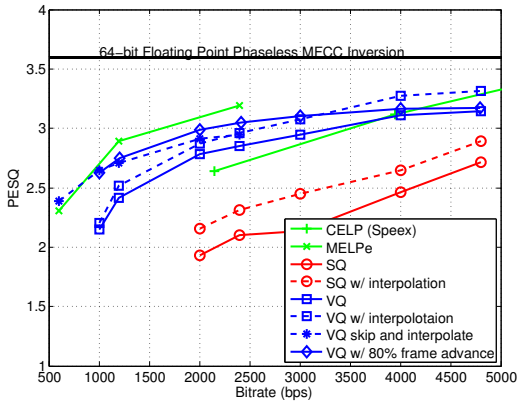


Fig. 4. PESQ scores for various MFCC coding schemes and other low bit-rate codecs.

C. Hybrid Scalar/Vector Quantization of MFCCs

We next consider a quantizer which utilizes both SQ and vector quantization (VQ) of the MFCCs. For a given target bit-rate, we utilize the non-uniform SQ described above but reserve 6 bits per frame. MFCCs that were allocated 1 bit per coefficient are grouped into 8-tuples and MFCCs that were allocated 2 bits per coefficient are grouped into pairs. These groups are then non-uniformly vector quantized while MFCCs that were allocated 3 or 4 bits per coefficient are scalar quantized as before. The 6 bits per frame previously reserved are used to quantize the MFCCs which were allocated 0 bits. It should be noted that the coefficient means are no longer required at the decoder, as no coefficients are allocated 0 bits.

This hybrid SQ/VQ codec improves performance as shown in Fig. 4 (blue square solid line). Again, there is a muffling associated with the reconstructed speech, but clarity is improved for all bit-rates. We can again insert interpolated frames to improve quality of the decoded speech as shown in Fig. 4 (blue square dashed line). Demonstrated by the insertion of interpolated frames, the window overlap has direct consequences for the quality of the quantized representation for a given bit-rate (less overlap means more bits available for each frame) as well as for the quality of the LSE-ISTFTM algorithm (more overlap increases the redundancy used by the LSE method). As a result, the amount of window overlap or frame advance was varied. At bit rates less then 3000 bps a frame advance of 80% with interpolation outperforms the standard 50% frame advance with interpolation, as ilustrated in Fig. 4 (blue diamond solid line). In fact, at the lowest bit rates (less then 1000 bps) we achieve the highest quality speech signals with *no* window overlap, shown in Fig. 4 (blue star dashed line). Thus it was empirically determined that for the lowest bit-rate (600 bps) it is better to decrease the window overlap and assign more bits to encode each MFCC vector and for higher bit-rates (1200, 2400, 4800 bps) it is better to increase the window overlap and reduce the number of bits for each MFCC.

D. MFCC-Based Codec Performance Comparison

The proposed codec is fully scalable to a wide range of bit-rates. The proposed MFCC-based codec was compared to other low bit-rate coding schemes, namely Code-Excited Linear Prediction (CELP) [17] and Mixed-Excitation Linear Predictive enhanced (MELPe) [13]–[16].

The CELP class of algorithms has been proven to work reliably and provide good scalability. Some examples of CELP-based standard codecs consist of G.728 [19] which operates at 16 kbps and DoD CELP (Federal Standard 1016) [20] which operates at 4.8 kbps. The open-source Speex codec, also based on CELP, operates at a variety of bit-rates ranging from 2150 bps to 44 kbps [21]. The MELPe algorithm was derived using several enhancements to the original MELP algorithm [13]. MELPe is also known as MIL-STD-3005 [14] and NATO STANAG-4591 [15] and supports bit-rates of 1200 bps and 2400 bps. There also exists a proprietary 600 bps MELPe vocoder algorithm [16].

The performance of CELP (Speex) and MELPe are shown in Fig. 4 (green lines) for various bit-rates between 600 and 4800 bps. The proposed MFCC-based codec yields PESQ scores better than the Speex codec for bit-rates ranging from 600 to 4800 bps. Additionally, the proposed MFCC-based codec matches performance of the state-of-the-art MELPe codec at 600 bps. Although speech files coded with the MELPe and Speex codecs are intelligible, they are hindered by the artificial, synthetic-sounding speech common to many formant based synthesis systems, especially when encoding at the each codec's minimum bit-rates. In contrast, the MFCC-based approach generates more natural sounding speech, but may contain subtle, raspy and scratchy artifacts.

V. Conclusions

In this paper, we have reviewed our previously-proposed method to reconstruct a speech frame from high-resolution, mel-frequency cepstral coefficients which relies on a pseudo-inverse of the mel-weighting functions and a phase estimate provided by the LSE-ISTFTM algorithm. Reconstruction of the speech waveform from MFCCs results in quality degradation of approximately one PESQ MOS point but nonetheless still leads to fair/good quality speech (~ 3.6 PESQ MOS). We have proposed a speech codec, based on a hybrid scalar/vector quantization of the MFCCs, which is scalable down to bit-rates as low as 600 bps. It was shown to have PESQ better than the CELP codec and matches the state-of-the-art MELPe codec at 600 bps. The proposed codec results in more natural sounding speech than those of existing codecs without the synthetic-sounding artifacts. Finally, use of an MFCC-based codec may facilitate speech processing algorithms which use MFCCs as features such as distributed speech recognition applications.

References

[1] T. F. Quatieri, *Discrete Time Speech Signal Processing*. Prentice Hall, 2002.

[2] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, and K. M. Indrebo, "Time-domain isolated phoneme classification using reconstructed phase spaces," *IEEE Trans. Audio, Speech, Language Process.*, vol. 13, no. 4, pp. 458–466, July 2005.

[3] J. Zeng and Z.-Q. Liu, "Type-2 fuzzy hidden Markov models and their application to speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 3, pp. 454–467, June 2006.

[4] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition–general classifier approaches and data fusion methods," *Pattern Recognition*, vol. 35, pp. 2801–2821, 2002.

[5] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[6] S. S. Stevens and J. Volkmann, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, Jan. 1937.

[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[8] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequenecy cepstral coefficients and pitch frequency," in *Proc. ICASSP*, vol. 3, 2000, pp. 1299–1302.

[9] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 24–33, 2007.

[10] L. E. Boucheron and P. L. D. Leon, "On the inversion of mel-frequency cepstral coefcients for speech enhancement applications," in *Proc. Int. Conf. Signals and Electronic Systems (ICSES)*, 2008.

[11] "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," European Telecommunications Standards Institute," ETSI ES 202 211 V1.1.1 (2003-11), 2003.

[12] X. Shao and B. Milner, "Clean speech reconstruction from noisy mel-frequency cepstral coefficients using a sinusoidal model," in *Proc. ICASSP*, vol. I, 2003, pp. 704–707.

[13] A. V. McCree and T. P. Bamwell, "Mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, July 1995.

[14] "Analog-to-digital conversion of voice by 2400 bits per second mixed excitation linear prediction," United States Military," US MIL-STD-3005, Dec. 1999.

[15] "The 1200 and 2400 blt/s nato interoperable narrow band voice coder," North Atlantic Treaty Organization," STANAG 4591 Ratification Draft 1, Dec. 1999.

[16] M. Chamberlain, "A 600 bps MELP vocoder for use on HF channels," in *Proc. IEEE Milcom Conference*, 2001.

[17] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1985, pp. 937–940.

[18] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[19] "Coding of speech at 16 kbit/s using low-delay code excited linear prediction," European Telecommunications Standards Institute," ITU - Recommendation G.728, Sept. 1992.

[20] J. Campbell, Jr., T. E. Tremain, and V. C. Welch., "The federal standard 1016 4800 bps CELP voice coder," *Digital Signal Processing*, vol. 1, no. 3, pp. 145–155, 1991.

[21] (2010). [Online]. Available: {http://speex.org}