Low Bit-Rate Speech Coding through Quantization of Mel-Frequency Cepstral Coefficients

Laura E. Boucheron, *Member, IEEE*, Phillip L. De Leon, *Senior Member, IEEE*, and Steven Sandoval

***Abstract*—In this paper, we propose a low bit-rate speech codec based on vector quantization (VQ) 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 phase information. By evaluating the contribution toward speech quality that individual MFCCs make and applying appropriate quantization, our results show that the MFCC-based codec exceeds the state-of-the-art MELPe codec across the entire range of 600–2400 bps, when evaluated with the perceptual evaluation of speech quality (PESQ) (ITU-T recommendation P.862). The main advantage of the proposed codec is in distributed speech recognition (DSR) since the MFCCs can be directly applied thus eliminating additional decode and feature extract stages; furthermore, the proposed codec better preserves the fidelity of MFCCs and better word accuracy rates as compared to CELP and MELPe codecs.**

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]–[5], and speaker characterization [6]–[8]. 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 [9], [10] is perhaps more common, e.g., [7], [8].

Reconstruction of the 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 [11]. 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. The authors invoke the work of [12] in which a bit rate of 4000 bps did not impair speech recognition rates. A technique was proposed for reconstructing the speech waveform for purposes of "playback" by either client or server. The technique relies on sinusoidal synthesis whereby the MFCCs along with pitch and a voicing decision allow sinusoidal amplitudes, frequencies, and phases to be estimated and used in reconstruction. The authors report that "natural sounding, good quality intelligible speech" can be reconstructed when 24 MFCCs per frame are used and pitch and voicing decision estimates are accurate [11].

The need to reconstruct speech from MFCCs gained further importance with an extension to the European Telecommunications Standards Institute (ETSI) Aurora distributed speech recognition (DSR) standard [13]. This extension includes a provision whereby a time-domain speech signal may be reconstructed from the received MFCC vectors (transmitted over a 4800 bps channel) together with fundamental frequency and voicing information (transmitted over a 800 bps auxillary channel) using sinusoidal synthesis [11], [14]. In potential applications of DSR, reconstruction of the speech waveform is important in cases of dispute arising from recognition errors or simply for human verification of transmitted utterances [15].

In [15] the authors investigate speech reconstruction solely from MFCC vectors. They estimate pitch and voicing from the MFCCs by exploiting correlation between the fundamental frequency and the spectral envelope. The primary result is a technique that can yield good predictions of pitch and voicing and, when coupled with MFCC vectors, enables speech reconstruction via sinusoidal synthesis similar to [11]. In their experiments, the authors use the ETSI Aurora standard of 13 MFCCs per vector extracted from a 25 ms speech frame at a rate of 100 vectors/s (uncoded). In informal listening tests, the authors report that "provided the fundamental frequency contour was smooth, then intelligible and reasonable quality speech can be reconstructed" [15]. Unfortunately, prediction accuracy of the fundamental frequency and voicing when using speaker independent models can be degraded. Therefore without formal subjective tests or objective quality measures, it is difficult to fully assess quality in the speech signal reconstructed from MFCCs through this approach.

In [16] the authors present a predictive vector quantization (VQ)-based coding scheme for 13 MFCCs at 8700 bps, nearly twice the ETSI bit-rate of 4800 bps. Speech reconstruction is accomplished by a conversion of MFCCs to linear prediction coefficients (LPCs) which are used to synthesize the speech waveform. The authors report PESQ scores equivalent to the

This work was presented in part at the Data Compression Conference (DCC), 2011.

L. Boucheron, P. De Leon, and S. Sandoval are with the Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces, NM 88003 USA. +1 (575) 646-3771 Tel, +1 (575) 646-1435 Fax, {lboucher, pdeleon, xxspsxx}@nmsu.edu

* Direct all correspondence regarding this manuscript to Phillip De Leon.

G.729 standard.

In contrast to previous work [11], [13]–[15], the method we present in this paper 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 work presented in this paper is an extension of that presented by the authors in [17]. Surprisingly, this approach does not appear to have been proposed despite yielding a much simpler reconstruction algorithm than the sinusoidal-synthesis based methods presented in [11], [13]–[15].

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 [15]. The key to this approach is to simply not discard too much information and instead use 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 [11], [13]–[15]. We show in this paper that through proper quantization of the MFCCs we can encode at 4800 bps rates (compatible with the ETSI Aurora DSR standard [13]) while at the same time enabling good quality, intelligible, reconstructed speech. Furthermore, we show that from the high-resolution MFCC vector we can easily downconvert to the low-resolution MFCC vector for compatibility with ASR. This conversion produces a low-resolution MFCC vector that is statistically closer to that which is directly extracted from a speech signal than other codecs based on LPCs.

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 Perceptual Evaluation of Speech Quality (PESQ) [18] scores better than the state-of-the-art Mixed-Excitation Linear Predictive enhanced (MELPe) codec [19]–[22] and at higher bit-rates, better than that of Code-Excited Linear Prediction (CELP)-based codecs [23]. Furthermore, in informal subjective listening tests, the proposed MFCC-based codec has a Mean Opinion Score (MOS) which closely follows PESQ.

This paper is organized as follows. In Section II, we present the procedure for reconstruction of the speech waveform from MFCCs. 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 in the decoder and present performance and computational benchmarking results in Section V. In Section VI, we present the procedure for downconversion of MFCC features for use in DSR systems and provide several measures of the accuracy of the downconverted MFCCs including Word Accuracy Rate (WAR). Finally, we conclude in Section VII.

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

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

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

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

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

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

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

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

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 MFCCs $\mathcal{M}$, we apply a set of weighting functions $\mathbf{\Phi}$ to the power spectrum prior to the Discrete Cosine Transform (DCT) and log operations [10]

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

This weighting $\mathbf{\Phi}$ is based on human perception of pitch [9] and is most commonly implemented in the form of a bank of filters each with a triangular frequency response [10]. 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) [10], 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 which is 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 as in [24]. 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-root 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 [25], for $|\mathbf{X}|^2$ as

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

2) Least-Squares Estimation of Speech Frame from the 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 therefore estimate the phase spectrum since this is discarded during computation of the MFCCs. The speech frame (and hence the phase spectrum) can be estimated using the well-known, Least-Squares Estimate, Inverse Short-Time Fourier Transform Magnitude (LSE-ISTFTM) algorithm shown in Algorithm 1 [26]. 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.

Algorithm 1 Least-Squares Estimate (LSE), Inverse Short-Time Fourier Transform Magnitude (ISTFTM).

Given: Speech magnitude spectrum, $|\hat{\mathbf{X}}|$ and desired number of iterations, M

Find: LSE speech frame, $\tilde{\mathbf{x}}$

1: Initialize $\tilde{\mathbf{x}}$ to white noise
2: **for** $m = 1$ **to** M **do**
3: Compute $\mathcal{F}\{\tilde{\mathbf{x}}\} = \tilde{\mathbf{X}} = |\tilde{\mathbf{X}}|e^{j\angle\tilde{\mathbf{X}}}$
4: Let $\hat{\mathbf{X}} = |\hat{\mathbf{X}}|e^{j\angle\tilde{\mathbf{X}}}$
5: Compute $\tilde{\mathbf{x}} = \mathcal{F}^{-1}\left\{\hat{\mathbf{X}}\right\}$
6: **end for**

III. Artifactual Effects in Reconstruction of Speech from MFCCs

In this section, we quantify the loss in speech quality when reconstructing a speech signal from MFCCs using the PESQ measure [18]. We will first provide a brief discussion of the PESQ measure and then discuss the degradation in the reconstruction of speech due to the two above mentioned sources of information loss: least-squares inversion of the mel-scale weighting functions and least-squares estimation of the speech frame from the magnitude spectrum.

A. Perceptual Evaluation of Speech Quality (PESQ)

PESQ is an objective measure of speech quality developed to obtain the highest correlation with subjective MOS [27] and was adopted by ITU-T as Recommendation P.862 [28]. PESQ is considered to have "acceptable accuracy" for evaluation of waveform-based (PCM) and CELP-based codecs for bit-rates $\geq$4000 bps [28]. Applications for which PESQ has not been officially validated include CELP-based codecs less than 4000 bps and non CELP-based codecs [28]. However, many researchers have found PESQ useful in evaluating codec performance outside of recommendation [18], [29]–[31]. PESQ has recently been evaluated against WAR for ASR and artificial voices [29]. In [29], the authors compare PESQ and other objective measures of speech quality to WAR for a digit recognition task measured for different noise conditions and noise-reduction algorithms. They find PESQ to be well-correlated to WAR, although the specific relationship is not linear. A least-squares fit to the PESQ-WAR relationship provides a means to estimate WAR from PESQ; this estimator was found to have coefficient of determination $R^2 = 0.85$. Thus, while PESQ may not have been specifically analyzed for speech quality evaluation of non-CELP-based codecs or low bit-rate codecs, it has been shown to be a broadly applicable measure across a wide range of speech processing applications. Furthermore, we present results of informal subjective listening tests and provide examples of speech waveforms decoded with our and other codecs in order to provide informal subjective validation of the PESQ results presented in this paper [32].

B. Quality of Speech Reconstructed from MFCCs

Although the DCT, log, and square root operations in (5) are all invertible, we must 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. 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 $\sim$ 24s in duration [33]. The baseline PESQ score for the TIMIT reference signals is 4.5.

We begin by computing the 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

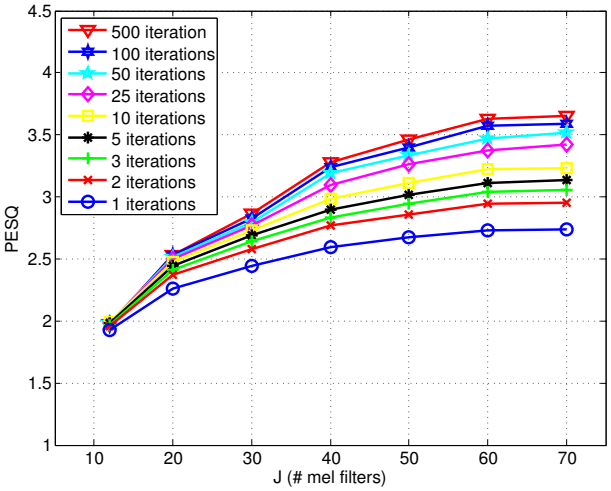


Fig. 1. Inversion of MFCCs in a clean phase-less environment at a various number of iterations. These results are averaged for a sample of 16 TIMIT speakers. We find that, for a large number of MFCCS ($J \geq 40$), 100 iterations provides a good balance between inversion quality and computation and yields a PESQ score within $\sim 2\%$ of the converged solution.

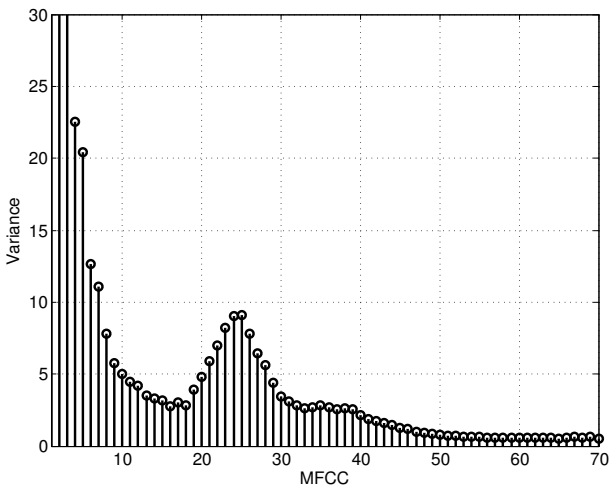


Fig. 2. Variance of individual MFCCs.

(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) and 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 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. We will thus use the LSE-ISTFTM algorithm with 100 iterations for all work and when evaluating the MFCC-based codec, a PESQ of 3.58 as our benchmark.

IV. Mel-Frequency Cepstrum-Based Speech Codec

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

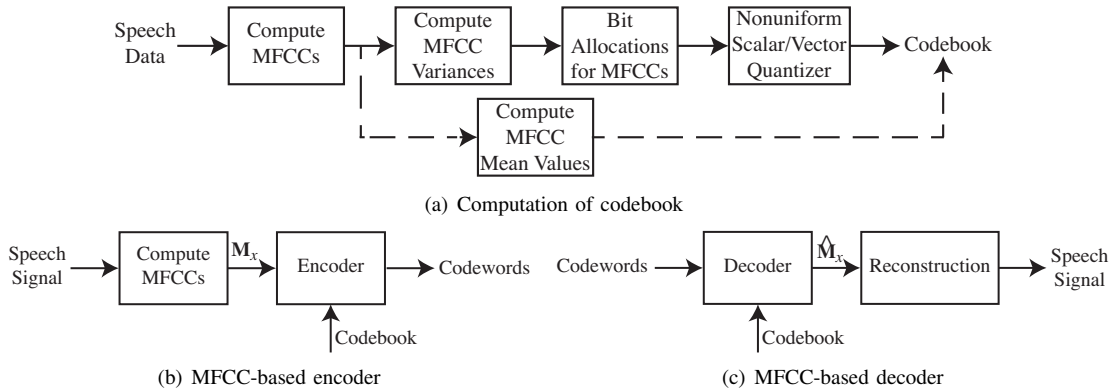
A. Assessing the Variance of Individual MFCCs as a Measure of Contribution to Speech Quality

We have observed through simulation and PESQ evaluation that the individual MFCC variance is directly related to its contribution to speech quality. Shown in Fig. 2 is a plot of the variance of individual MFCCs across the speech frames for the complete 630 speaker TIMIT corpus. We see a large variance for the first seven MFCCs and smaller variance 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 *high-resolution* MFCCs to formant structure and correspondence of middle coefficients to pitch period (i.e., vocal excitation) information. In the high-resolution MFC (70 coefficients), we can roughly partition coefficients 2-14 as corresponding to formant structure and coefficients 15-30 as corresponding to pitch; the first MFCC is related to energy. It is the high-resolution aspect of these MFCCs that allows for direct modeling of the pitch information by the MFC. Since the initial 1 kHz of the spectrum is binned with a one-to-one correspondence, the harmonic structure of the pitch is maintained in the spectrum and converted by the DCT to the middle portion of the cepstrum. Fig. 2 thus suggests that more bits will have to be allocated to the few MFCCs corresponding to formant structure.

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], [34]. 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 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



(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 the LSE-ISTFTM algorithm.

j is

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

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) in which the blocks summarize the above information. In summary, 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 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.

C. Varying the Window Overlap

As an analysis tool, MFCCs are normally computed using overlapped windows in order to minimize edge effects. A typical value is 50% overlap as used in the SQ codec. We have found, however, that window overlaps other than 50% can allow for further improvement of the quality of low bit-rate speech signals. 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).

Empirically, it was 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. In fact, at the lowest bit rate we achieve the highest quality speech signals with *no* window overlap. In the following section we will utilize various window overlaps to achieve highest performance.

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 (magenta 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-B).

D. Vector Quantization of MFCCs

We next consider split VQ compression of the MFCCs. Experiments have shown that VQ produces superior performance over SQ even when the components of the input vector are statistically independent [35]. We were limited by computation and memory when determining more than 2^{14} VQ points, using fast k-means [34]. Thus we split the 70-dimensional MFCC vector into subvectors each coded with no more than 14 bits each. The number of subvectors is determined by the bit-rate (higher bit rates allowed more bits per frame and hence more subvectors). Training data was composed of 200 speakers from the TIMIT corpus, each recording was approximately 6 seconds in length.

As a result, we encode MFCC vectors at different bit-rates as specified in Table I. For example, at 1200 bps with a 25% window overlap, a total of 27 bits are available to encode

TABLE I

ALLOCATION OF BITS FOR VQ CODEC AT DIFFERENT BIT-RATES. BIT-RATE SPECIFIES THE TARGET BIT-RATE, OVERLAP IS THE PERCENTAGE OVERLAP USED IN ANALYSIS, AND BITS/FRAME IS THE TOTAL NUMBER OF BITS AVAILABLE TO CODE EACH RESULTING FRAME. ENERGY, FORMANT, AND PITCH SPECIFY THE CODING UTILIZED FOR EACH SUBSET OF MFCCS. # INTERP. FRAMES SPECIFIES THE NUMBER OF INTERPOLATED FRAMES INSERTED PRIOR TO RECONSTRUCTION, AND EQUIVALENT OVERLAP IS THE RESULTING EQUIVALENT OVERLAP FOR THE LSE-ISTFTM ALGORITHM.

Bit-rate	Overlap	Bits/ Frame	Energy (Coeff 1)	Formant (Coeffs 2–14)	Pitch (Coeffs 15–70)	# Interp. Frames	Equivalent Overlap
600 bps	0%	18	4-bit SQ	14-bit VQ		7	87.5%
1200 bps	25%	27	4-bit SQ	14-bit VQ	9-bit VQ	3	81.25%
2400 bps	25%	54	4-bit SQ	14-bit VQ (Coeffs 2–6) 14-bit VQ (Coeffs 7–14)	14-bit VQ (Coeffs 15–30) 8-bit VQ (Coeffs 31–70)	3	81.25%
4800 bps	25%	108	4-bit SQ	14-bit VQ (Coeffs 2–4) 14-bit VQ (Coeffs 5–7) 14-bit VQ (Coeffs 8–10) 14-bit VQ (Coeffs 11–14)	14-bit VQ (Coeffs 15–22) 14-bit VQ (Coeffs 23–30) 10-bit VQ (Coeffs 31–50) 10-bit VQ (Coeffs 51–70)	3	81.25%

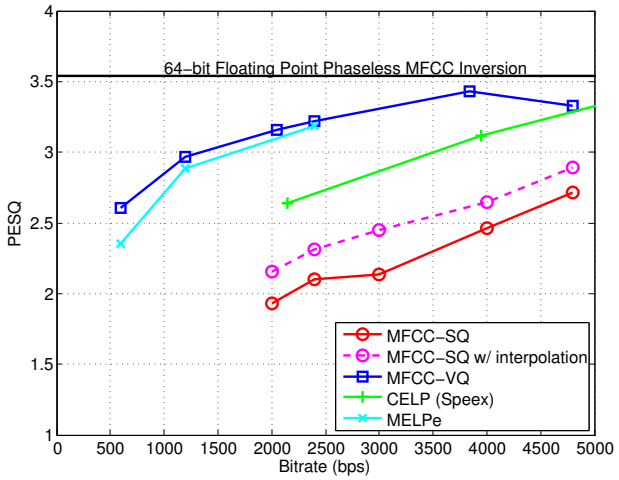


Fig. 4. PESQ scores for various MFCC coding schemes and other low bit-rate codecs. These results are averaged for the same 16 TIMIT speakers.

each MFCC vector. The 1st coefficient, which is related to energy and has the highest variance of any MFCC, is encoded using 4 bit SQ, coefficients 2-14 which contain formant information are encoded using 14 bit VQ, and coefficients 15-70 which contain pitch information are encoded using 9 bit VQ. Interpolated frames are inserted prior to the reconstruction of the waveform with the LSE-ISTFTM algorithm; overlap, bit allocation, and interpolation parameters are listed in Table I along with the equivalent reconstruction overlap. It should be noted that the coefficient means (shown in Fig. 4) are not required for the VQ, as no coefficients are allocated 0 bits as in the SQ coder.

The performance of the VQ codec is significantly improved over the SQ codec, especially at low bit-rates as shown in Fig. 4 (blue square line). Additionally, we can achieve PESQ scores above 2.5 for bit-rates as low as 600 bps. Again, there is a muffling associated with the reconstructed speech, but clarity is improved over the SQ-based MFCC codec for all bit rates.

V. RESULTS

A. Objective Performance as Measured by PESQ

The proposed MFCC-based codec was compared to other low bit-rate coding schemes, namely MELPe [19]–[22] and CELP [23]. It is important to note that the proposed MFCC-based codec is based on completely different theory than MELP and CELP codecs. Rather than modeling the speech production process with linear prediction, we are quantizing the MFCCs themselves and then directly reconstructing speech from the MFCCs.

The CELP class of algorithms has been proven to work reliably as well as provide good scalability. Some examples of CELP-based standard codecs consist of G.728 [36] which operates at 16 kbps and DoD CELP (Federal Standard 1016) [37] 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 [38]. There are several aspects of human speech that cannot be modeled by traditional LPCs. This results in a difference between the original and reconstructed speech, which is referred to as the residue. CELP-based codecs first compute the LPCs and then calculate the residue. The residue is compared to a code book and the code word which best represents the residue is transmitted. A synthesis filter in the decoder utilizes the residue for more accurate synthesis.

The MELPe algorithm was derived using several enhancements to the original MELP algorithm [19]. MELPe is also known as MIL-STD-3005 [20] and NATO STANAG-4591 [21] and supports bit-rates of 600, 1200, and 2400 bps. Traditional LPC algorithms use either periodic pulse trains or white noise as excitation for a synthesis filter. The MELPe family of vocoders use a mixed-excitation model of the human voice and extensive lookup tables to extract and regenerate speech. The MELPe codec also utilizes aperiodic pulse excitation, pulse dispersion to soften the synthetic sound of reconstructed speech, adaptive spectral filtering to model the poles of the vocal tract, and multiple stage vector quantization (MSVQ) for compression. The MELPe codec is further tuned to code the English language. In this work, we utilize software by Compandent, Inc. for MELPe transcoding.

The performance of CELP (Speex) and MELPe are shown in Fig. 4 (green plus and cyan x lines, respectively) for various bit-rates between 600 and 4800 bps. The proposed MFCC-based codec yields PESQ scores better than the CELP and MELPe codecs for bit-rates ranging from 600 to 4800 bps. Although the decoded speech files coded with the CELP and MELPe codecs are intelligible, they are hindered by the artificial, synthetic-sounding speech common to many formant

TABLE II
SUBJECTIVE LISTENING TEST RESULTS. PESQ REFERS TO THE OBJECTIVE SCORE AS SHOWN IN FIG. 4 AND MOS REFERS TO THE AVERAGE MEAN OPINION SCORE OF THE 20 SUBJECTIVE TEST PARTICIPANTS. % DIFF IS THE PERCENTAGE DIFFERENCE, COMPUTED WITH RESPECT TO PESQ; THUS, NEGATIVE VALUES INDICATE A DEVIATION WITH PESQ SCORING LOWER THAN MOS AND VICE VERSA.

Codec	Bitrate	PESQ	MOS	% Diff
MFCC	600 bps	2.60	2.75	-5.8
	1200 bps	2.97	2.83	4.9
	2048 bps	3.16	3.00	5.1
	2400 bps	3.22	3.13	3.0
	3840 bps	3.43	3.33	3.1
MELPe	600 bps	2.36	2.68	-13.4
	1200 bps	2.89	3.50	-21.1
	2400 bps	3.18	3.85	-21.1
CELP	2150 bps	2.64	2.75	-4.2
	3950 bps	3.12	3.08	1.4

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 contains raspy and scratchy artifacts. We have posted example audio files on the Internet for interested readers [32].

It should be noted that the ETSI DSR standard [13] targets a 4800 bps codec, with the possibility of an additional 800 bps for auxiliary information (5600 bps total). It is clear from these results, that while the proposed MFCC codec requires a higher resolution MFC than proposed by ETSI, we can still achieve a bit-rate well within the ETSI allocation and still reconstruct reasonable quality speech. To the best of the authors' knowledge, there are no published results for quality of the reconstructed speech using the ETSI proposed back-end speech reconstruction algorithm [13]. It should be noted, however, that the ETSI standard [13] bit-rate includes error-control coding, while all other codecs discussed in this section do not.

B. Subjective Performance as Measured with Listening Tests

We recruited 20 native English speakers (10 males and 10 females) to participate in an informal subjective listening test of the proposed codec. Subjects were asked to rate audio files based on the standard five point Mean Opinion Score (MOS) rating scale [18]. Each subject was presented with two transcoded TIMIT files (one male and one female) for the MELPe, CELP, and MFCC VQ codecs at each of the bitrates plotted in Fig. 4; the TIMIT files for each condition were chosen at random.

Table II shows the results for the subjective listening tests, along with the objective results using the PESQ metric for comparison. We find a good match between PESQ and MOS scores for both the proposed MFCC codec and the CELP codec; the percentage difference for these codecs is less than 6%. This validates the use of PESQ for our non CELP-based codec. For the MELPe codec, however, we find much larger deviation between PESQ and MOS, ranging from 13% to 21%. We hypothesize that subtle quality enhancements due to postfiltering (and other proprietary improvements) may have abnormally influenced the test subjects with respect to the MELPe codec.

TABLE III
COMPUTATION TIME FOR THE PROPOSED ENCODER AND DECODER FOR ∼24 SECOND SPEECH FILES.

Bitrate	Encoder	Decoder
600 bps	3.0 s	14.9 s
1200 bps	2.1 s	10.4 s
2400 bps	4.7 s	10.4 s
4800 bps	8.7 s	10.8 s

We thus find the proposed codec to be better in PESQ and competitive in an informal MOS rating than the CELP and MELPe codecs. The proposed MFCC codec, however, has additional advantages when considered for DSR or ASR systems, as will be discussed in further detail in Section VI.

C. Post-Processing in the MELPe Codec

Several codecs take advantage of post processing techniques to achieve higher quality speech. MELPe for example, uses both adaptive spectral filtering and a pulse dispersion filter. The adaptive spectral filtering is designed to sharpen formant peaks in voiced speech and the pulse dispersion filter is designed to spread the signal decay in unvoiced speech. These post filtering processes are specifically designed for LPC codec artifacts and improve quality and reduce the harshness of the synthesized speech. Future work on the MFCC-based codec should explore the use of other post processing techniques to improve speech quality of the proposed codec.

D. Benchmarking for the Proposed MFCC Codec

The encoding process involves two stages: MFCC computation and quantization. MFCC feature vectors are very common in the speech processing realm and do not pose a computational concern even when high resolution MFCCs are extracted. The quantization stage is the main computational concern in the encoder. A non-optimized MATLAB® implementation of the encoder was benchmarked on a standard 3.4 GHz PC; Table III displays the resultant average encoder times for the ∼24 second speech files. We see that even the worst case scenario at 4800 bps is approximately 3 times faster than real-time.

The decoding process involves two stages: decoding and feature inversion. The decoding process is simply a lookup operation and does not pose a computational concern. Feature inversion is accomplished by means of the iterative LSE-ISTFTM algorithm. The majority of the complexity of LSE-ISTFTM algorithm is due to a forward and inverse STFT every iteration. A non-optimized MATLAB® implementation of the decoder was similarly benchmarked; Table III displays the resultant average decoder times. The worst case average total decoder time occurs at 600 bps because the effective overlap after interpolated frames are inserted is greatest (87.5% vs. 81.25%). However, the worst case scenario at 600 bps is still approximately 2 times faster than real-time.

E. Scaling Complexity for the Proposed MFCC Codec

Given better-than-real-time benchmarks in MATLAB® for encoder and decoder, a C/C++ compiled code executing on

a modern microprocessor or DSP is unlikely to present an unreasonable burden on a mobile device or a speech server for DSR applications. However, there are at least two ways to decrease complexity and execution time without significant degradation in performance.

The bulk of the complexity in the encoder resides in the distance calculations between each data point and every codeword. Methods such as hierarchical VQ or codeword clustering have been proposed address this problem and can be effectively employed to significantly reduce this aspect of complexity. Although not currently implemented for this work, such methods could significantly reduce encoder complexity with little to no effect on speech quality. We refer the reader to [35] for more information.

The bulk of the decoder complexity resides in the LSE-ISTFTM algorithm, and most of the LSE-ISTFTM complexity resides in the FFTs computed in each iteration (see Algorithm 1). We can decrease the number of FFTs in two ways: by omitting interpolated frames inserted at the decoder (upsampling), or by decreasing the number of iterations. Reducing the upsampling by a factor of (2, 4) computation time drops by approximately a factor of (2, 4) and PESQ decreases by ($<$1%, $<$15%). Reducing the number of iterations by a factor of (2, 4, 10), PESQ decreases by ($<$2%, $<$10%, $<$11%). By reducing the upsampling by a factor of 4 and reducing iterations by a factor of 10, computation time reduces by a factor of 40, and PESQ drops by $<$18%.

VI. Application to Distributed Speech Recognition

A. Downconversion of High Resolution MFCCs

In most speech processing applications such as ASR which use MFCCs as features, e.g., the widely used Hidden Markov Model Toolkit (HTK) [39], the number of MFCCs extracted per speech frame is between 13 and 20. This is the motivation behind the establishment of a standard MFCC-based data channel for DSR [13].

As has been discussed, our proposed MFCC-based codec requires a much higher resolution mel-frequency cepstrum (more MFCCs per speech frame) in order to obtain good-quality speech reconstruction. However, downconversion from a high-resolution MFCC vector to a low-resolution MFCC vector for feature compatibility with ASR systems is relatively simple. The implementation of this conversion can be seen by reviewing (5). To recover the power spectrum, we need only invert the DCT and log operations and left multiply by pseudoinverse of the mel weighting matrix, $\mathbf{\Phi}^{\dagger}$. We then apply a new, lower-dimension mel weighting matrix $\mathbf{\Phi}$ ($K \times L$ where $K < J$) followed by the log and DCT operations. It is only the pseudo-inverse that introduces any error; thus, we can quantify the error due to downconversion by considering the downconverted mel filters.

We have investigated the effect of this downconversion process by considering the error, $|\mathbf{\Phi}_{24} - \mathbf{\Phi}'_{24}|$, where $\mathbf{\Phi}_{24}$ are the standard 24 mel-scale weighting functions, $\mathbf{\Phi}'_{24} = \mathbf{\Phi}_{24}\mathbf{\Phi}^{\dagger}_{70}\mathbf{\Phi}_{70}$ are the downconverted mel-scale weighting functions, and $\mathbf{\Phi}_{70}$ are the high-resolution mel-scale weighting functions used in this work. We find that in the downconversion process, the 24 mel-scale weighting functions are modified by less than 10% of their original values when downconverted from a high resolution MFC.

B. Accuracy of Downconverted MFCCs in ASR

Standard ASR systems use Gaussian Mixture Models (GMMs) of MFCCs (features) extracted from short segments of speech in order to recognize phonemes or other units of a word. Therefore, one way to assess the impact on ASR accuracy when using downconverted MFCCs is to statistically measure the deviation between reference MFCCs (those extracted from a reference or uncoded signal) and the downconverted MFCCs. By focusing on the features themselves, we avoid having to control many of the other variables that impact ASR accuracy. In addition, this deviation can bring insight into the impact of MELPe and CELP coding when features are extracted from transcoded signals and applied to ASR. In particular, we are interested in evaluating the fidelity of MFCCs under coding.

In this work, we propose to use the Kullback-Leibler (KL) divergence between a GMM of reference MFCCs and a GMM of downconverted MFCCs. Although KL divergence is a widely-used measure, there is no known closed-form solution for KL divergence between GMMs. However, as described in [40], KL divergence between GMMs can be approximated as

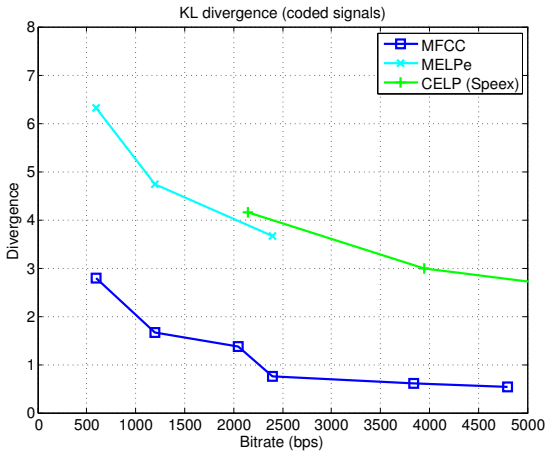
$$\mathrm{KL}(\lambda_c||\lambda_r) \approx \frac{1}{M}\sum_{m=1}^{M}\log p\left(\mathcal{M}_{c,m}|\lambda_c\right) - \frac{1}{M}\sum_{m=1}^{M}\log p\left(\mathcal{M}_{c,m}|\lambda_r\right) \quad (9)$$

where $\mathcal{M}_{c,m}$ is the m-th vector of downconverted MFCCs from the proposed codec, λ_c is the GMM of $\mathcal{M}_{c,m}$, and λ_r is the GMM of MFCCs extracted from the reference speech signal. Alternately, we can use a symmetric version of (9) as described in [41]

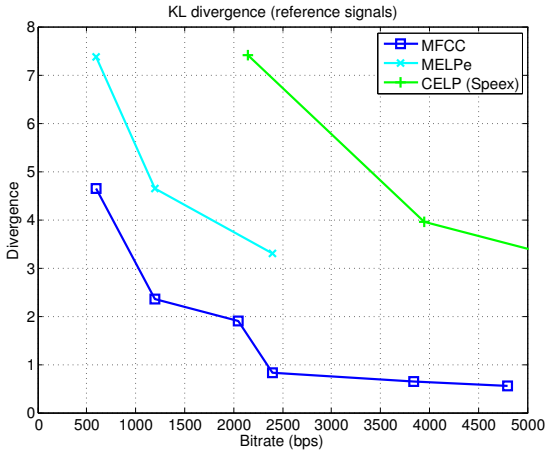
$$\mathrm{KL}(\lambda_r||\lambda_c) \approx \frac{1}{M}\sum_{m=1}^{M}\log p\left(\mathcal{M}_{r,m}|\lambda_r\right) - \frac{1}{M}\sum_{m=1}^{M}\log p\left(\mathcal{M}_{r,m}|\lambda_c\right) \quad (10)$$

where $\mathcal{M}_{r,m}$ is the m-th vector of reference (uncoded) MFCCs.

According to the measures in (9) and (10), preservation of the MFCCs will result in small KL divergence. Figure 5 displays KL divergence for (9) and (10) where $\mathcal{M}_{c,m}$ is either the downconverted MFCCs from the proposed codec or the MFCCs extracted from the MELPe or CELP codecs averaged across all 16 TIMIT-sample speakers. For this figure, we used 32 component densities for the formulation of the GMMs λ_c and λ_r, and 13 MFCCs (similar to the ETSI Aurora standard) for each vector. The proposed codec achieves significantly lower KL divergence than either CELP or MELPe across the range of bitrates. The proposed codec therefore preserves



(a) KL divergence using coded signals per (9).



(b) KL divergence (symmetric) using reference signals per (10).

Fig. 5. Kullback-Leibler divergence values for different codecs. For each of the codecs considered, the divergence decreases as the bitrate increases due to the higher audio fidelity. Note that the proposed MFCC-based codec displays significantly lower divergence than either MELPe or CELP.

the fidelity of the MFCCs better than MELPe or CELP and is expected to achieve ASR accuracies closer to reference (uncoded) signals than MELPe and CELP codecs.

C. Automatic Speech Recognition

We have implemented a large vocabulary, speaker independent ASR system using HTK, trained on the entire Wall Street Journal 0 (WSJ0) corpus [42] using word-internal triphones. We used 80 test files from the November 1992 ARPA WSJ test set [42].

Table IV presents the Word Accuracy Rate (WAR) for each of the codecs at a variety of bitrates. We find that the proposed MFCC-based codec achieves higher WARs than all codecs at all bit rates and at 2400 bps achieves WAR comparable to the uncoded speech signals. The WARs closely mirror the results indicated by the KL divergence plots as shown in Fig. 5. In addition, these WARs closely mirror the PESQ results shown in Fig. 4, consistent with [29].

TABLE IV
WORD ACCURACY RATE (WAR) FOR THE PROPOSED MFCC CODEC, MELPE, AND CELP ON 80 WSJ0 FILES.

Codec	Bitrate	WAR
Uncoded	64000 bps	88.2%
MFCC	600 bps	76.6%
	1200 bps	86.6%
	2400 bps	88.7%
MELPe	600 bps	73.5%
	1200 bps	82.9%
	2400 bps	86.0%
CELP	2150 bps	74.7%
	3950 bps	79.4%

VII. CONCLUSIONS

In this paper we have proposed a method for low bit-rate coding of speech through vector quantization of the mel-frequency cepstral coefficients. Reconstruction of the speech waveform from the MFCCs is accomplished by a least-squares inversion of the mel weightings and a least-squares estimate of the inverse STFT. The proposed codec is scalable down to bit-rates as low as 600 bps and was shown to have better speech quality as measured by PESQ than the CELP and MELPe codecs. In informal subjective listening tests, the proposed MFCC-based codec has a mean opinion score which closely follows PESQ results. Additionally, we have shown that the proposed codec better preserves the fidelity of MFCCs, used as features in DSR, and leads to better word accuracy rates than either MELPe or CELP codecs.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Oded Gottesman of Compandent Inc. for use of the MELPe encoder which was used to prepare MELPe transcoded speech signals for comparison. The authors would also like to thank Dr. Alan McCree of the MIT Lincoln Laboratories for suggestions on our research and for reviewing an early draft of this manuscript. The authors gratefully acknowledge Mr. Don McCoy of the Embedded Systems Division, Mentor Graphics for assistance in the configuration, training, and testing of the ASR.

REFERENCES

[1] T. F. Quatieri, *Discrete Time Speech Signal Processing*. Prentice Hall, 2002.
[2] M. D. Skowronski and J. G. Harris, "Increased MFCC filter bandwidth for noise-robust phoneme recognition," in *Proc. ICASSP*, vol. I, 2002, pp. 801–804.
[3] 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, Jul. 2005.
[4] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, vol. I, 2006, pp. 325–328.
[5] 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, Jun. 2006.
[6] 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.
[7] 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.

[8] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173–192, 1995.

[9] 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.

[10] 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.

[11] 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.

[12] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments," in *Proc. ICASSP*, 1998, pp. 977–980.

[13] "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.

[14] 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.

[15] 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.

[16] G. H. Lee, J. S. Yoon, and H. K. Kim, "A MFCC-based CELP coder for server-based speech recognition in network environments," in *Proc. European Conference on Speech Communication and Technology (INTERSPEECH)*, 2005, pp. 1369–3172.

[17] L. E. Boucheron, P. L. D. Leon, and S. Sandoval, "Hybrid scalar/vector quantization of mel-frequency cepstral coefficients for low bit-rate coding of speech," in *Proc. Data Compression Conference*, 2011, pp. 103–112.

[18] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC, 2007.

[19] 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, Jul. 1995.

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

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

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

[23] 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.

[24] L. E. Boucheron and P. L. D. Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications," in *Proc. IEEE Int. Conf. on Signals and Electronic Systems (ICSES)*, 2008, pp. 485–488.

[25] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[26] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[28] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, Telecommunication Standardization Sector, ITU-T Recommendation P.862, 2001.

[29] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artifical voice," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2006–2013, Nov. 2006.

[30] 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.

[31] K. Harwell, G. Scheets, J. Weber, and K. Teague, "A multilanguage study of the quality of interleaved MELP voice traffic over a lossy network," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 565–568, Jul. 2009.

[32] (2011). [Online]. Available: "{http://www.ece.nmsu.edu/~pdeleon/Research/MFCC_Coding_Demo.zip}"

[33] (2011). [Online]. Available: {http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/timit.zip}

[34] C. Elkan, "Using the triangle inequality to accelerate k-means," in *Proc. Int. Conf. Machine Learning (ICML)*, 2003.

[35] A. Gersho and R. M. .Gray, *Vector Quantization & Signal Compression*. Kluwer Academic Publishers, 1992.

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

[37] 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.

[38] (2011). [Online]. Available: {http://speex.org}

[39] (2011). [Online]. Available: {http://htk.eng.cam.ac.uk}

[40] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unscented transform and its application to speaker recognition," in *Proc. INTERSPEECH*, 2005, pp. 1985–1988.

[41] M. Ben, R. Blouet, and F. Bimbot, "A Monte Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002.

[42] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," http://www.keithv.com/pub/baselinewsj, Cavendish Laboratory, University of Cambridge, Tech. Rep., 2006.

Laura E. Boucheron (S'01-M'07) received the B.S. and M.S. degrees in Electrical Engineering from New Mexico State University in 2001 and 2003, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of California Santa Barbara in 2008. She has intern and graduate research experience at both Sandia National Laboratories and Los Alamos National Laboratory and postdoctoral and research faculty experience in the Klipsch School of Electrical and Computer Engineering at New Mexico State University. She is currently Assistant Professor in the Klipsch School.

Phillip L. De Leon (SM'03) received the B.S. Electrical Engineering and the B.A. in Mathematics from the University of Texas at Austin, in 1989 and 1990 respectively and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Colorado at Boulder, in 1992 and 1995 respectively. Currently, he is a Professor in the Klipsch School of Electrical and Computer Engineering and Director of the Advanced Speech and Audio Processing Laboratory at New Mexico State University.

Steven Sandoval received the B.S. Electrical Engineering in 2007, and his M.S. Electrical Engineering in 2010 from the Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces. He is presently working on his Ph.D. degree in electrical engineering at New Mexico State University and also works as a system analyst for a department of defense contractor.