# Enhancing Distributed Speech Recognition with Back-End Speech Reconstruction

*Tenkasi Ramabadran, Jeff Meunier, Mark Jasiuk, and Bill Kushner*

Speech Processing Research Laboratory
Motorola Labs
Schaumburg, IL 60196, USA
{ramabadr,meunier,jasiuk,kushner}@labs.mot.com

## Abstract

In this paper, we present a method to enhance the usefulness of a Distributed Speech Recognition (DSR) system by providing it the capability to reconstruct speech at the back-end. Speech reconstruction is achieved using the standard DSR parameters, viz., Mel-Frequency Cepstral Coefficients (MFCC) and log-energy, and some additional parameters, viz., voicing class, pitch period, and (optionally) higher-resolution energy information. From the MFCC parameters and energy information, the spectral magnitudes at the harmonics of the pitch frequency are estimated. Based on the class information, the harmonic phases are appropriately modeled. The harmonic magnitudes and phases are used to reconstruct speech according to the well-known sinusoidal model for speech synthesis [4][5]. Transmission of the additional parameters for speech reconstruction increases the DSR bit rate by less than 20%. Evaluation by Mean-Opinion-Score (MOS) test and Diagnostic Rhyme Test (DRT) show that speech reconstructed as above is of reasonable quality and quite intelligible.

## 1. Introduction

The performance of an automatic speech recognition system using speech transmitted over a mobile channel as input can be significantly degraded when compared to a system that uses the original unmodified speech as input [1][2]. This degradation is due to the distortions introduced in the transmitted speech by the coding algorithm used as well as to channel transmission errors. A Distributed Speech Recognition (DSR) system overcomes this problem by separating the *feature extraction* part and the *pattern matching* part of a speech recognizer as follows. A mobile unit at the *front-end* extracts appropriate recognition features from the original un-coded speech and transmits them over an error-protected data channel to the *back-end* server located at some point on the network for completing the recognition operation. With this approach, there are no coding distortions and the transmission channel has very little effect on the recognition system performance. Moreover, the mobile unit has to perform only the computationally inexpensive feature extraction part thereby leaving the more complex pattern matching part to the back-end server. The European Telecommunications Standards Institute (ETSI) has recently published a standard [3] for the front-end feature extraction and compression algorithms for use in a DSR system. The features extracted are *Mel-Frequency Cepstral Coefficients* (MFCC) and *log-energy* for each speech frame. The extracted features are compressed, error-protected, and formatted into a bit stream for transmission at a rate of 4800 bps. The standard also describes suitable algorithms for bit stream decoding and channel error mitigation.

While a DSR system offers many advantages for implementation of a speech-recognition-based application over a mobile channel, it has the drawback that the original speech (in coded form) is not available at the back-end for storage and/or verification purposes. The availability of the original speech at the back-end is desirable for several reasons including:

- Enabling applications that require human assistance, e.g., dictation systems
- Storage and verification of legally sensitive information, e.g., financial transactions
- Validation of utterances during database collection over the DSR channel for training and system tune-up

An obvious solution to providing the original speech at the back-end is to employ a separate speech coder but at the cost of significantly increased bandwidth. A bandwidth-efficient solution would necessarily have to use the DSR parameters, viz., MFCC and log-energy, as part of the speech coder information. Such an approach is currently under serious consideration within ETSI. In this paper, we describe a solution based on a low data rate (2400 bps) parametric vocoder developed within Motorola. This vocoder is a harmonic (or sinusoidal) vocoder [4][5] and uses the following parameters: *voicing class*, *pitch period*, *sub-frame energies*, and *harmonic magnitudes* modeled using line spectral frequencies (LSF). Since the MFCC parameters provide a reasonable representation of the speech spectrum from which the harmonic magnitudes can be estimated, they can replace the LSF. Similarly, the log-energy parameter can be used by itself or with additional higher-resolution energy information to replace the sub-frame energies. The additional parameters needed for reconstruction of speech at the back-end increases the DSR bit rate by only 10% to 20%. Subjective evaluation tests show that speech reconstructed by the method presented in this paper is of reasonable quality and quite intelligible.

Speech reconstruction using the MFCC parameters has recently been considered by other authors [6]-[9]. The method presented here is similar to that of Dan Chazan, et. al. [6][7] but differs in the technique used to estimate the speech harmonic magnitudes from MFCC and also in several details of analysis and synthesis.

In section 2, we describe the technique used to estimate the harmonic magnitudes from MFCC and compare it with the technique presented in [6]. Details of analysis to extract the additional parameters needed for reconstruction as well as the

synthesis procedure are discussed in section 3. In section 4, results of subjective evaluation tests, viz., MOS test and DRT, are presented. Section 5 concludes the paper and provides some directions for future research.

## 2. Estimation of Harmonic Magnitudes

Before we describe the technique for estimating the speech harmonic magnitudes from MFCC, it is useful to look at how the MFCC features are extracted [3] in the first place. Digitized speech, e.g., at 8000 samples per second (Fs) and 16 bits/sample, is first passed through a DC-offset removal filter and divided into overlapping frames, each 200 samples long. At this point, the frame energy is computed at and its (natural) logarithm is used as the log-energy parameter. A pre-emphasis filter is then used to emphasize the higher frequency components. Next, each speech frame is (Hamming) windowed and transformed into the frequency domain by means of a 256-point FFT (Fast Fourier Transform). The FFT magnitudes in the frequency range between 64 Hz and 4000 Hz ($F_s/2$) are then Mel-filtered as follows. First, the frequency range above is warped into a Mel-frequency scale using the expression

$$Mel(f) = 2595.0 * \log_{10}(1 + \frac{f}{700.0}). \quad (1)$$

The Mel-frequencies corresponding to 64 Hz and 4000 Hz are respectively 98.6 and 2146.1. This range is then divided into 23 equal-sized, half-overlapping *bands* (or *channels* or *bins*) with each band 170.6 wide and the center of each band 85.3 apart. The center of the first band is located at 98.6+85.3 = 183.9, and that of the last band is located at 2146.1-85.3 = 2060.8. These bands of equal size in the Mel-frequency domain correspond to bands of unequal sizes in the linear frequency domain with the size increasing along the frequency axis. The FFT magnitudes falling inside each band are then averaged (filtered) using a triangular weighting window (with the weight at the center equal to 1.0 and at either end equal to 0.0). The Mel-filter bank outputs are then subjected to a (natural) logarithm operation. The 23 log-spectral values are then transformed into the cepstral domain by means of a 23-point DCT (Discrete Cosine Transform) to obtain the 13 MFCC values C0 through C12. The values C13 through C22 are discarded, i.e., not computed. The 13 MFCC values are then quantized and transmitted to the back-end. The MFCC and log-energy values are updated every 10 ms in the DSR standard.

To estimate the speech harmonic magnitudes from the 13 MFCC values, we proceed as follows. First, we perform an IDCT (Inverse DCT) to transform the MFCC values back into the Mel-frequency domain. A 23-point IDCT of the MFCC values C0 through C12 (assuming C13 through C22 are zeros) would restore the original 23 log-spectral values except for the distortion caused by truncation of the cepstral sequence and quantization. These log-spectral values correspond to the centers of the 23 frequency bands. However, we need the log-spectral values at other frequencies also and to accomplish this we increase the resolution of the IDCT by a factor of (2K+1) where, K>0. For example, if K=85, the higher resolution IDCT introduces 170 additional Mel-frequency points between the band centers and 85 additional points before the first and after the last band center. The size of the IDCT in this case is 23*171=3933, and its resolution is 85.3/171= 0.4988 in the Mel-frequency scale. The first and last Mel-frequency points are respectively 183.9-85*0.4988 = 141.5 and 2060.8 + 85*0.4988 = 2103.2. These correspond to the linear frequencies of 93.6 Hz and 3824.6 Hz respectively. Log-spectral values outside this range can be assumed to be equal to that of the nearest available frequency point without any serious effect. The higher resolution IDCT essentially interpolates between the frequency band centers using the DCT basis functions themselves as the interpolating functions. To facilitate computation of the IDCT at selected frequency points, an IDCT matrix L of size 12x3933 is pre-computed and stored as follows.

$$L_{i,j} = \left(\frac{2}{23}\right)\cos\left(\frac{(2j+1)*i*\pi}{2*23*171}\right),$$

i = 1, 2, …, 12; j = 0, 1, …, 3932. (2)

The zeroth row corresponding to C0 is implicit and need not be stored since its value is constant at 1/23 for all columns. The rows corresponding to C13 through C22 need not be stored as these coefficients are unavailable and assumed to be zero. To get the log-spectral value at any given Mel-frequency, we find the nearest frequency point and the corresponding column of L and form an inner product with the MFCC vector [C0, C1, …, C12].

From the log-spectral value, the spectral magnitude can be computed by exponentiation. However, the effect of the pre-emphasis filter has to be removed from this spectral magnitude. Computing the magnitude response of the pre-emphasis filter at the desired frequency and dividing the spectral magnitude easily accomplishes this. Besides the pre-emphasis filter, the Mel-filter also emphasizes higher frequencies because of the increasing width of the frequency bands along the linear frequency axis. The Mel-filter magnitude response at any band center can be taken to be the width of the corresponding band, and for any other frequency, we can use an interpolated value. In fact, we can compute a combined magnitude response of the pre-emphasis filter and the Mel-filter so that the effect of both filters can be removed from the spectral magnitude in a single step. An elegant alternative is to compute the MFCC values corresponding to the pre-emphasis filter impulse response and subtract these values from the speech MFCC values.

The following steps summarize the technique for estimating the harmonic magnitudes from the MFCC values:

1. From the speech MFCC values, subtract the (fixed) MFCC values corresponding to the impulse response of the pre-emphasis filter.
2. For each harmonic frequency, warp the frequency into Mel-scale and find the nearest frequency point and the corresponding column of L. Form an inner product of this column and the modified MFCC vector from step 1 to compute the log-spectral value for the harmonic.
3. Exponentiate the log-spectral value to get the spectral magnitude for the harmonic.

We now compare the technique presented above with the one outlined in [6] where, the harmonic magnitudes are obtained by sampling a linear combination of frequency-domain basis functions. The non-negative basis function gains are determined using an iterative procedure such that the Mel-filter bank outputs of the reconstructed speech is similar to the Mel-filter bank outputs obtained from the original MFCC by IDCT and exponentiation. To do the comparison, we

computed the distortion (in dB) between the original harmonic magnitudes and the estimated harmonic magnitudes using the two techniques. For this purpose, a speech database, sampled at 8 kHz, pre-processed with a modified IRS filter, and consisting of 32 sentence pairs (4 males + 4 females, 4 sentence pairs each) was used. The original harmonic magnitudes were obtained by analyzing each voiced frame (20 ms long) for the pitch period and the FFT magnitudes corresponding to the pitch harmonics. The Motorola speech vocoder analysis program was used for this purpose. The DSR front-end program [3] was then used to compute the MFCC vectors making sure that the two procedures were time-aligned. The technique presented in this paper as well as the one in [6] was then used to estimate the harmonic magnitudes for each 20ms frame from (every second) MFCC vector. The average distortion D over N voiced frames was computed as

$$D = \frac{1}{N}\sum_{n=1}^{n=N} Dn \qquad (3)$$

where, the distortion for the n$^{th}$ frame is given by

$$Dn = \sqrt{\frac{1}{Ki}\sum_{k=1}^{k=Ki}\left[20*\log_{10}(M_{k,n}) - 20*\log_{10}(\tilde{M}_{k,n})\right]^2} \qquad (4)$$

In (4), K is the number of harmonics, and $M_{k,n}$ & $\tilde{M}_{k,n}$ are the original and estimated harmonic magnitudes respectively. For each frame, the original and estimated magnitudes were first normalized such that their log-mean is zero. The results of the comparison are shown in Table 1. In addition to comparing the two techniques, the effect of quantization and truncation are also presented in this table. For informational purposes, the modeling distortion corresponding to the 14 (unquantized) LSF values used in the speech vocoder is only 2.63 dB.

**Table 1**. Comparison of Estimation Techniques

| Input parameters used | Distortion in dB (N = 4768) | |
|---|---|---|
| | Technique from [6] | Technique from this paper |
| 13 MFCC values (quantized) | 5.03 | 4.64 |
| 13 MFCC values (unquantized) | 4.79 | 4.33 |
| 23 MFCC values (unquantized) | 4.19 | 3.80 |

## 3. Speech Vocoder Details

In this section, we briefly describe the Motorola speech vocoder and give details of how it was modified to provide reconstructed speech at the back-end of a DSR system.

### 3.1. Analysis

The speech vocoder operates on 22.5 ms frames and has a bit rate of 2400 (54 bits/frame). For each frame, it extracts the following parameters: class, pitch period, sub-frame energies, and harmonic magnitudes modeled by line-spectral frequencies. The class parameter is quantized using 2 bits and indicates whether a frame is non-speech, unvoiced speech, mixed-voiced speech or voiced speech. Speech/Non-speech classification is done using an energy-based Voice Activity Detector (VAD). Determination of voicing level is based on a number of features including periodic correlation (normalized correlation at a lag equal to a pitch period), aperiodic energy ratio (ratio of energies of decorrelated and original frames), and high-frequency energy ratio. The pitch period has a range between 19 and 139 samples and is (non-uniformly) quantized using 7 bits. It is estimated using a time-domain correlation analysis of low-pass filtered speech. The pitch period of non-speech and unvoiced speech frames is set at 139. The sub-frame energies are quantized in the log-domain using 8 bits by a 4-dimensional VQ. For non-speech and unvoiced speech frames, the energy is computed over a sub-frame (4 sub-frames per frame). For voiced frames, it is computed over a pitch period. The harmonic magnitudes are estimated from a 256-point FFT of Hamming windowed speech at the pitch harmonics. They are modeled using a 14$^{th}$ order all-pole model and the model parameters, viz., LSF, are quantized using a 4-split VQ with 37 bits.

In modifying the vocoder to work with a DSR system, the frame size was reduced to 20 ms, and the class, pitch period, and sub-frame energies were transmitted as additional parameters along with the DSR bit stream. This corresponds an additional 17 bits/frame or 850 bps. In a second version, the sub-frame energies were not transmitted. This corresponds to an increase in bit rate of only 450 bps. At the back-end, the harmonic magnitudes were estimated from (every second) MFCC vector for both versions. For the second version, the sub-frame energies were computed from the DSR log-energy parameter through interpolation.

### 3.2. Synthesis

The synthesizer reconstructs speech using a sinusoidal model of speech production:

$$s(j) = \sum_k A_{k,j}\cos(\Phi_{k,j}) \qquad (5)$$

where, the speech sample *s(j)* is synthesized as the sum of a number of harmonically related sinusoids with amplitude $A$ and phase $\Phi$. In (5), $j$ is the sample index and $k$ is the harmonic index. From the transmitted parameters, the synthesizer computes the number of harmonics, the amplitudes, and the phases for each sample index $j$ so that *s(j)* can be computed. These values are first computed at the midpoint of each frame as follows. The pitch frequency and its harmonics are computed from the pitch period. Using the LSF (or MFCC for DSR back-end speech reconstruction) and the sub-frame energy corresponding to the midpoint of the frame, the harmonic magnitudes are estimated. The computation of harmonic phases depends on the class parameter. For voiced speech, the phases are computed using a phase model derived from a real-life excitation (inverse-filtered speech) domain pulse. To these, the phases corresponding to the spectral envelope (which are obtained from an all-pole model of the harmonic magnitudes) are added. For non-speech and unvoiced speech frames, randomly chosen phases are used. In the case of mixed-voiced frames, voiced speech model is used for lower frequencies and unvoiced speech model is used for higher frequencies. From the modeled phases, any linear phase component (corresponding to a time-shift) is removed.

Once the amplitudes at the midpoints of the current and previous (voiced) frames are known, the amplitudes at the sub-

frame boundaries are computed using linear interpolation and adjusted for the energies at these points. Amplitudes within a sub-frame are computed using linear interpolation. The harmonic phases at different sample indices are computed by allowing the phases to evolve (linearly) according to the frequency. These frequencies are allowed to change at the sub-frame boundaries in equal steps from the previous values to the current values. Any phase discontinuities arising out of this evolution are resolved using linear phase correction factors (slight frequency shifts). If the previous and current frames are of different classes (e.g., one is voiced and the other is unvoiced) or both are voiced but the pitch periods are quite different, e.g., doubling, the two frames are synthesized independently and overlap-added in the time-domain.

## 4. Subjective Evaluation

To evaluate the quality of both versions of DSR back-end speech synthesizers, a subjective Mean-Opinion-Score (MOS) test was performed. The same speech database that was used to compare the estimation techniques (Section 2) was used for this purpose. A total of 32 conditions were included in the test. Several MNRU conditions [10] and coding standards were included to serve as references and to ensure that the entire range of quality level was spanned. A group of 32 naïve listeners was used to evaluate the speech quality based on a 5-point scale: Bad (1), Poor (2), Fair (3), Good (4), and Excellent (5). The test was conducted in a soundproof room and the speech samples were presented through a headphone mono-aurally.

The MOS numbers (averaged over 256 votes) for some important conditions are as follows. The original un-coded speech scored 4.32. The G726 (32 Kbps ADPCM) [11] and the G729 (8 Kbps CS-ACELP) [12] standards scored respectively 3.65 and 3.89. The MELP vocoder (2400 bps Federal standard) [13] scored 2.93. The 2400 bps Motorola vocoder and its 20 ms variation at 2700 bps scored 3.11 and 3.15 respectively. The first version of the DSR back-end synthesizer (850 bps additional) scored 2.43. The second version (450 bps additional) scored 2.26.

Besides quality, intelligibility is quite important for DSR back-end speech synthesis. To evaluate intelligibility, a (limited) Diagnostic Rhyme Test (DRT) [14] was performed. Standard DRT test words spoken by 2 speakers (1 male and 1 female) and 8 (untrained) listeners were used in the test. The test was conducted in a soundproof room and the speech samples were presented through a headphone mono-aurally.

The (averaged) overall intelligibility score for the 20 ms version of the Motorola vocoder (2700 bps) was found to be 88. The score for the first version of the DSR back-end synthesizer (850 bps additional) was found to be 82.8.

## 5. Conclusions

In this paper, we described a method for speech reconstruction at the back-end of a Distributed Speech Recognition system. The method uses the standard DSR parameters, viz., Mel-Frequency Cepstral Coefficients and log-energy, and some additional parameters, viz., voicing class, pitch period, and (optionally) higher-resolution energy information for the reconstruction. The additional parameters increase the DSR bit rate by less than 20%. Subjective evaluation tests show that the reconstructed speech is of reasonable quality and good intelligibility.

In future work, we intend to combine the DSR log-energy parameter with the sub-frame energies of the vocoder to reduce the bit rate. We expect to achieve some quality improvement by estimating the harmonic magnitudes from the MFCC vectors every 10 ms instead of every 20 ms. Use of noise suppression to improve performance under noisy conditions and channel coding to protect against transmission errors are also being investigated.

## 6. References


[1] S. Euler and J. Zinke, "The Influence of Speech Coding Algorithms on Automatic Speech Recognition," *Proceedings of Int. Conf. on Acoust., Speech, and Signal Processing* (ICASSP), pp. 621-624, 1994.

[2] B.T. Lilly and K.K. Paliwal, "Effect of Speech Coders on Speech Recognition Performance," *International Conference on Spoken Language Processing* (ICSLP), pp. 2344-2347, 1996.

[3] European Telecommunications Standards Institute, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," *ETSI Standard* ES 201 108 v1.1.2, April 2000.

[4] R.J. McAulay and T.F. Quatieri, "Speech Analysis / Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 34, pp. 744-754, August 1986.

[5] R.J. McAulay and T.F. Quatieri, "Sinusoidal Coding," in *Speech Coding and Synthesis* (W.B. Kleijn and K.K. Paliwal, eds.), Ch. 4, pp. 121-170, *Elsevier*, 1995.

[6] Dan Chazan, Ron Hoory, Gilad Cohen and Meir Zibulski, "Speech Reconstruction from Mel Frequency Cepstral Coefficients and Pitch Frequency," *Proceedings of Int. Conf. on Acoust., Speech, and Signal Processing* (ICASSP), pp. 1299-1302, 2000.

[7] Dan Chazan, Gilad Cohen, Ron Hoory and Meir Zibulski, "Low Bit Rate Speech Compression for Playback in Speech Recognition Systems," *European Signal Processing Conf.* (EUSIPCO), pp. 1281-1284, 2000.

[8] Z. Tychtl and J. Psutka, "Speech Production Based on the Mel Frequency Cepstral Coefficients," 6th *European Conference on Speech Communication and Technology* (EUROSPEECH), pp. 2335-2338, 1999.

[9] Z. Tychtl and J. Psutka, "Pitch Synchronous Residual Excited Speech Reconstruction on the MFCC," *European Signal Processing Conf.* (EUSIPCO), pp. 761-764, 2000.

[10] ITU Recommendation P.810, "Modulated Noise Reference Unit," 1996.

[11] ITU Recommendation G.726, "40, 32, 24, 16 Kbps Adaptive Differential Pulse Code Modulation (ADPCM)," 1990.

[12] ITU Recommendation G.729, "8 Kbps CS-ACELP Speech Coder," 1996.

[13] L.M. Supplee, R.P. Cohn, J.S. Collura and A.V.McCree, "MELP: The New Federal Standard at 2400 bps," *Proceedings of Int. Conf. on Acoust., Speech, and Signal Processing* (ICASSP), pp. 1591-1594, 1997.

[14] ANSI Standard S3.2, "Method for Measuring the Intelligibility of Speech over Communication Systems," 1989.