Speech Coding Using Trajectory Compression and Multiple Sensors

Sorin Dusan, James Flanagan, Amod Karve, and Mridul Balaraman

Center for Advanced Information Processing (CAIP)
Rutgers University, Piscataway, NJ, U.S.A.
{sdusan,jlf,amod,mridulb}@caip.rutgers.edu

Abstract

This paper presents a new method of multi-frame speech coding based upon polynomial approximation of speech feature trajectories incorporating multiple sensor signals from microphones, accelerometer, electro-glottograph, and microphone. The trajectory polynomial approximation exploits the inter-frame information redundancy encountered in natural speech. The trajectory method is applicable to features such as spectral parameters, gain, and pitch. The method is suitable for application to a frame vocoder to further reduce the transmission bit rate. Multiple transducers increase the intelligibility and quality of the coded speech in noisy environments. Experimental results are obtained by embedding the new method into an enhanced mixed-excitation linear prediction vocoder. The resulting vocoder operates at 1533 bps and preliminary intelligibility and quality tests show results comparable to those of the original 2400 bps vocoder.

1. Introduction

Frame vocoders analyze speech with sequential short-time windows and encode parameters that characterize the spectrum, gain, pitch, and voicing in each of these frames. Frame windows usually range in duration between 20 ms and 50 ms and are overlapped, resulting in frame steps of 10 ms to 30 ms. Longer window durations average speech features across phonological boundaries and larger window steps do not resolve short speech sounds. Various frame vocoders achieve transmission bit rates ranging from 2400 bps to 16000 bps.

A popular coding technique for feature vectors in frame vocoders is vector quantization [1]. Feature vectors are matched with prototype vectors which are stored in a codebook and the index of the closest codebook vector is transmitted by the encoder. A variation of this algorithm is the multi-stage vector quantization (MSVQ) in which feature vectors are quantized using a number of successive codebooks and stages and the bits of the resulting indexes are concatenated and transmitted by the encoder.

An efficient frame vocoder algorithm at low bit rate is the mixed-excitation linear prediction (MELP) algorithm [2], which is currently the U.S. Federal vocoder standard at 2400 bps. In this standard the frame window duration is 25 ms and the frame step is 22.5 ms. Each frame contains 54 bits and there are two types of frames: voiced and unvoiced. Depending on the type of frame, the vocoder encodes parameters such as line spectral frequencies (LSF), Fourier magnitudes, gain, pitch, overall voicing, bandpass voicing, and aperiodic flag. The quantization of the LSF parameters is accomplished by a MSVQ algorithm consisting of four stages.

A synchronization bit is transmitted in each frame and in unvoiced frames 13 bits are used for error protection.

Frame vocoders achieve low bit rate transmission with relatively high quality and intelligibility. Although the frame step (or frame rate) is optimized according to the rate of speech sounds occurring in natural speech, successive speech frames still exhibit information redundancies in the speech features. To further compress the signal and reduce its transmission rate a number of N successive frames can be analyzed and encoded together. This aims at eliminating some of the information redundancies common to the successive frames. There are various ways to achieve this and to encode the multi-frame parameters. One popular technique uses matrix quantization, in which N feature vectors, each of dimension M, from successive frames are jointly quantized as an MxN matrix [3]. A multi-frame variant of the MELP algorithm is the enhanced mixed excitation linear prediction (MELPe) vocoder [4]. This vocoder contains a noise preprocessor and operates at either 1200 bps or 2400 bps. At 1200 bps the vocoder jointly encodes the speech parameters of three successive frames. The LSF parameters of the three frames are quantized using a forward-backward interpolation method.

In this paper we present an approach to multi-frame coding in which polynomial functions of order P are estimated from N successive frames and are employed to derive P+1 feature vectors to be encoded and transmitted instead of the original N feature vectors. Compression of the feature parameters is achieved if P+1<n. The trajectory compression by polynomial approximation can be applied to various speech features such as linear prediction parameters (LPC), line spectrum frequencies (LSF), gain parameters, and pitch. By choosing appropriate P and N values, various compression rates can be achieved.

To improve the quality and intelligibility of coded speech in noisy environments, experimental results are presented in this paper by combining input speech signals obtained from multiple transducers including microphones, accelerometer, electro-glottograph, and electromagnetic micropower sensor. The microphones and the accelerometer are used to provide a combined input signal which is more robust to noise, whereas all the sensors are used to estimate more robust pitch and overall voicing parameters.

In this paper the trajectory compression technique is combined with the multi-sensor inputs this combination is embedded in a MELPe implementation. Experimental results are presented to demonstrate performance. Subjective listening tests for three different noisy environments are reported for the new vocoder operating at 1533 bps.

The outline of the paper is as follows. Section 2 presents the principles of the trajectory compression method based on
polymer approximation. Section 3 describes the multi-sensor processing for noisy environments. Section 4 presents experimental results by embedding the combined methods into a MELPe implementation. Conclusions are presented in Section 5.

2. Trajectory Compression by Polynomial Approximation

Normal speech typically contains on average 10 to 15 sounds/s. This means that the average duration across sounds (or phonemes) is approximately 67 ms to 100 ms, depending on the speech rate (which is also a factor of speaking style, language, dialect, etc.). However, the frame step must be shorter than 67-100 ms because of two reasons: first, there exist speech sounds much shorter than the average sound duration; second, for high quality and intelligibility the encoder must not only extract and transmit the acoustic characteristics at the central (stable) position of these speech sounds but also for the transitions into and from these positions. These constraints lead to frame steps of about 10 to 30 ms. But these lower values cause the successive frames to be correlated and display information redundancies. A way to further compress the information contained in successive frames is to model the trend of speech features in segments containing a number of successive frames, as proposed in [5]. Thus, instead of encoding and transmitting the features from a number of successive frames, the trend functions of these features can be encoded and transmitted. This can result in fewer transmitted bits if compact approximation functions are employed. Then, the decoder can reconstruct the features of the original number of successive frames from the trend functions.

Compression of information can be achieved if the trend functions can be encoded with fewer bits than the bits used to encode the original segment of successive frames. Polynomial functions can satisfy the above requirement if

\[ P + 1 < N, \]  

where \( P \) represents the order of the polynomial functions and \( N \) is the number of frames in the multi-frame segment. Each individual feature trajectory \( FT_N \) represented by \( N \) feature values can be approximated in the least squares sense by a polynomial function \( f_P(t) \) of order \( P \) in accordance with

\[ FT_N = [s_1, s_2, ..., s_{N-1}, s_N], \]  

\[ f_P(t) = a_1 t^P + a_2 t^{P-1} + ... + a_P t + a_0, \]  

where \( s_1, s_2, ..., s_{N-1}, s_N \) are the \( N \) speech feature values in the trajectory, \( a_1, a_2, ..., a_P, a_0 \) are the polynomial coefficients, and \( t \) is the time variable representing the frame index within the multi-frame segment. For example, if the LSF feature vectors contain 10 features (vector elements), there are 10 unique polynomial functions order \( P \), each approximating the trajectory of an individual feature. The choice of \( P \) and \( N \) values affects the compression ratio. On one hand, lower \( P \) values and higher \( N \) values give higher compression ratios. On the other hand, too low \( P \) values can increase too much the approximation error and too high \( N \) values can lead to a too large algorithmic delay in the communication channel. Speech spectral features generally exhibit smooth transitions between adjacent sounds. The LSF parameters are known for their reduced dynamics and smooth behavior and in coding they are preferred over the linear prediction coding (LPC) parameters. The polynomial functions can accurately approximate, in the least squares sense, these feature trajectories with an arbitrary error that goes to zero when

\[ P + 1 = N. \]  

In this special case the polynomial functions do not approximate the feature trajectories but interpolate their frame values. Other features such as pitch and gain can also be modeled by polynomial functions, although the approximation errors for the same \( P \) and \( N \) values might be larger due to greater dynamic behavior.

Once the polynomial functions have been fit to the feature trajectories using a selected pair of \( P \) and \( N \) values, these functions are used to compute \( P+1 \) new values \( f_P(t_i) \) for each individual feature at arbitrary time positions within the multi-frame segment \( 1 \leq t_1 < t_2 < ... < t_{P+1} \leq N \). These new values are considered the feature values of \( P+1 \) new frames and they are encoded and transmitted instead of the original \( N \) frames of the segment. The originally developed codebook can be used to transmit the feature vectors representing the new \( P+1 \) frames values derived from the polynomial functions. At the reception, these \( P+1 \) values are used to compute a unique polynomial function of order \( P \) for each individual feature. Then, from these new polynomial functions the original feature values for the \( N \) frames are approximated by equation 3.

A feature compression factor \( C_F \) describing the percentage of transmitted bits saved compared to the originally transmitted bits for the \( N \) frames is given by

\[ C_F = \frac{1 - (P + 1)/N}{100\%}. \]  

For example, polynomial functions of order \( P=2 \) (quadratic functions) used to compress a speech segment containing \( N=6 \) frames can yield a saving in bits of \( C_F = 50\% \) comparing to the bits necessary to transmit the un-compressed feature trajectories. However, vocoders also transmit other bits in addition to those necessary to represent the continuous speech features and thus the total compression factor \( C_T \) will be less than 50% in this example. Selecting appropriate \( P \) and \( N \) values requires optimizing the following criteria: high feature compression factor, low trajectory approximation errors (high subjective intelligibility and quality scores), and low algorithm delay (low communication latency). The class of polynomial functions proves to provide good approximations of the speech feature trajectories and allows a substantial saving in the transmitted bits.

3. Multi-Sensor Processing

The quality and intelligibility of the coded speech in noisy environments decrease because some of the speech parameters to be encoded are affected by noise. One way of improving these parameters is to use multiple transducers to capture the acoustic signal and some physiological measures of speech such as the glottal excitation function. This paper...
reports on two methods using multiple sensors to improve speech quality and intelligibility in noisy environments.

3.1. Fusion of multiple input signals

The first speech improvement method investigated in this study was originally proposed in [6] and is based on fusing the signals from two or three microphones: a resident microphone placed at approximately 25 mm from the speaker’s mouth, a physiological microphone placed on the speaker’s throat above the glottis, and a bone conduction microphone placed on the speaker’s head. The fusion method, as described in [6], combines two signals: the physiological microphone signal after it is low-pass filtered at 300 Hz; the resident microphone signal after it is high-pass filtered at 300 Hz. The weight of the first signal is selected at 0.3 whereas the one of the second signal is selected at 1.0. Another version of this fusion method combines three signals: the physiological microphone signal after it is low-pass filtered at 200 Hz, the bone conduction microphone signal after it is band-pass filtered between 200 Hz and 700 Hz, and the resident microphone signal after it is high-pass filtered at 700 Hz. The weight of the first signal is selected at 0.5 whereas the ones of the second and third signals are selected at 1.0. These fusion frequencies and weights were experimentally determined by the authors of [6]. The effect of these fusion methods is an improvement in the quality and intelligibility of coded speech in noisy environments. These improvements are demonstrated by measurements presented in section 4.

3.2. Detection of pitch and voicing from multiple signals

The second method for improving the coded speech in this paper uses multiple input signals to extract the pitch and overall voicing parameters from speakers in noisy environments. Three types of sensors are used in this study: a resident microphone as described in the previous subsection, an electro-glottograph (EGG) sensor placed on the speaker’s neck around the glottal region, and a glottal electromagnetic micropower sensor (GEMS) placed on the speaker’s throat, as described in [7]. The GEMS sensor provides two similar signals that indirectly describe the glottal excitation function, whereas the EGG sensor provides a single signal whose waveform approximates the glottal opening function. An autocorrelation method as described in [2] and [7] is employed to estimate the pitch and voicing parameters from each of these signals. Then a median voting method is used to select the coded and transmitted values for these two parameters. Results are presented in the next section.

4. Experimental Results

The techniques for trajectory compression and for multiple speech sensors were combined and embedded into a 2400 bps MELP vocoder implementation [4]. Experiments of speech coding using segments of \(N=10\) frames and various polynomial orders \(P\) between 1 and 5 were carried out. Three types of features were compressed: LSF parameters (10), gain parameters (2), and pitch (1). These experiments simulated speech transmission at various bit rates less then 2400 bps, as a result of various feature compression factors ranging between 80% and 40%, respectively. In this section quantitative and qualitative results are reported for segments containing \(N=10\) frames (22.5 ms frame step) and polynomial functions of order \(P=4\). These values result in a feature compression factor \(C_f = 50\%\), and an equivalent transmission bit rate of 1533 bps.

Fig. 1 presents an example of the original LSF trajectories (dotted) and polynomial approximated functions (solid) for a segment of speech of 10 frames from a typical sentence from the TIMIT database. At the top of the figure the corresponding TIMIT phonetic transcriptions and frame numbers are also displayed.

![Figure 1: LSF trajectories of \(N=10\) frames approximated by polynomial functions of order \(P=4\) (\(C_f = 50\%)\).](image)

Fig. 2 shows two sets of spectra (left-original) and (right-that reconstructed from the \(C_f = 50\%)\) compressed trajectories for a whole TIMIT sentence. On the left side of each spectral set the corresponding TIMIT phonetic transcriptions and frame numbers are also displayed.

![Figure 2: Original and reconstructed from compressed trajectories LPC spectra for a TIMIT sentence (\(C_f = 50\%)\).](image)
high noise environments: Blackhawk Helicopter (BH), M2 Bradley Fighting Vehicle (M2), and Military Operations on Urbanized Terrain (MOUT). The average sound pressure level (SPL) in the three environments was, 110 dBC, 114 dBC, and 113 dBC respectively. Two types of utterances were processed from each speaker in each environment: diagnostic rhyme test (DRT) utterances containing 232 words and Harvard phonetically balanced sentences (HPBS). Four quantitative distortion measures were used to evaluate the compression method: the Itakura-Saito distance (IS) and the root mean squared errors for spectral distortion (RMS-S), gain (RMS-G), and pitch (RMS-P). Table 1 shows the distortion results averaged across speakers and environments computed between the original features and the reconstructed features with a 50% feature compression factor (equivalent of 1533 bps).

Table 1: Distortion measures for $C_f = 50\%$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRT</td>
<td>0.08</td>
<td>1.57</td>
<td>1.03</td>
<td>2.11</td>
</tr>
<tr>
<td>HPBS</td>
<td>0.13</td>
<td>1.88</td>
<td>1.33</td>
<td>2.18</td>
</tr>
</tbody>
</table>

Fig. 3 shows spectrograms of the original input noisy signal from the M2 environment and coded signals by MELPe at 2400 bps and by the new vocoder at 1533 bps.

Table 3 presents DRT intelligibility scores, averaged across speakers in each environment. In each environment the new vocoder at 1533 bps outperforms MELPe at 2400 bps.

Table 3: DRT intelligibility scores

<table>
<thead>
<tr>
<th>Noise environment</th>
<th>MELPe at 2400 bps</th>
<th>Vocoder at 1533 bps</th>
</tr>
</thead>
<tbody>
<tr>
<td>BH</td>
<td>79.25</td>
<td>81.97</td>
</tr>
<tr>
<td>M2</td>
<td>71.90</td>
<td>78.21</td>
</tr>
<tr>
<td>MOUT</td>
<td>84.33</td>
<td>86.26</td>
</tr>
</tbody>
</table>

5. Conclusions

This report implements and evaluates a method of compressing feature trajectories by polynomial approximation combined with a method for improving noise robustness based on multiple sensors. The evaluation uses quantitative distortion measures and subjective intelligibility tests. The experimental results for the new vocoder at 1533 bps show relatively low distortion measures and higher DRT scores than MELPe vocoder at 2400 bps in the same conditions.

6. Acknowledgements

We acknowledge Tom Quatieri of MIT Lincoln Laboratory for first proposing the approach of fusing the resident microphone and the physiological microphone output signals, and along with Kevin Brady, also from MIT Lincoln Laboratory, for providing fused files and Matlab scripts for fusion. We thank John D. Tardelli from ARCON Corp. for providing the DRT intelligibility scores. This research was supported by the Defense Advanced Research Projects Agency under the Advanced Speech Encoding program. *(Approved for Public Release, Distribution Unlimited).*

7. References