Proc. of NCC 2010, Chennai, India

Trajectory and surface modeling of LSF for low rate speech coding

M. Deepak and Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai-400076
{deepu, prao}@ee.iitb.ac.in

Abstract—**In low rate speech coders, frame-based speech spectral parameters, represented by line spectral frequencies (LSF), are typically encoded by vector quantization without exploiting explicitly the temporal correlation between frames. Recently, however, methods have been proposed that model the temporal trajectory of each LSF over a speech segment by a low-order polynomial function. In the present work, we propose a method that exploits the joint spectro-temporal correlation across a speech segment. The LSF values are treated as points on a smooth surface obtained by fitting a low-order bivariate polynomial function of the variables, frame index and LSF index. The issues that arise in incorporating such joint modeling in a low bit rate coder are identified, and some comparative results of performance are presented.**

I. INTRODUCTION

Speech coders aim at reducing bit rate while maintaining good quality of reconstructed speech. Most low bit rate coders are model based coders. Model based coders represent speech by a fixed set of parameters per frame, typically at a frame rate of 50 frames per second. This limits the achievable compression with parameter transmission needed at 20 ms intervals. Speech, however, is known to be relatively stationary over much longer intervals depending on the underlying phone class. This redundancy can be exploited by encoding larger segments of speech where speech parameters vary only slowly across frames thus facilitating more efficient quantization. In the present work, we investigate methods for segment level encoding of spectral parameters in the context of the Multiband Excitation (MBE) speech model.

The MBE model of speech is a powerful representation of the speech signal providing for compactness, perceptually superior quality and robustness to acoustic background noise [1]. Each 20 ms interval of speech is represented by a pitch, harmonic spectral amplitudes and voicing band decisions. The spectral amplitudes are further represented by a spectral envelope parameterized by LSFs. Efficient quantization of the LSFs is therefore crucial to achieving low bit rate coding while maintaining good reconstructed speech quality.

The LSF vector (typically 10 dimensional) is generated in every frame interval (20 ms). However, the temporal evolution of the LSFs across the frames is smooth leading to inter-frame redundancy. Modeling the temporal evolution of the LSFs by trajectory functions can help in exploiting the inter-frame redundancy. Polynomial functions have been used for modeling of pitch, LSFs and gain parameters of 2400 bps standard MELPe coder to achieve a rate of 1533 bps with nearly same speech quality as standard MELPe coder [2]. In [3], discrete cosine transform coefficients are used to represent the trajectory of individual LSFs and a bit rate reduction of nearly 50 % is achieved.

The above methods model the trajectory of each LSF independently. However it is well known that frame-level spectral parameters exhibit significant joint spectro-temporal correlation. Such correlation exists for LSFs as well as for decorrelated spectral representations such as MFCC when one considers time segments on the order of phone durations. The physiological basis for the spectro-temporal correlation is the nature of the articulatory dynamics. Joint modeling in the spectro-temporal domain could potentially contribute to further efficiencies in coding over the independent steps of coding of spectral coefficients (by frame-level VQ) and coding of temporal trajectories of the spectral coefficients. This new idea is proposed and developed in the present work.

Joint spectro-temporal modeling can be viewed as modeling a surface with a low-order polynomial over the time-frequency index plane. In the case of the MBE coder, the time index is the frame number while the frequency index is the LSF index. The value of the LSF (in Hz) is the surface variable whose dependence on the two indices is to be modeled. One method of representing a smooth surface is by a low order bivariate polynomial function. In this work, we explore the implementation of modeling the joint variation of LSFs by bivariate polynomial modeling and compare the performance in terms of quality of reconstructed speech with that obtained with temporal trajectory modeling only. The surface modeling is carried out over short homogenous segments of speech.

The organization of paper is as follows. Automatic segmentation of the speech signal is presented in Section II. The trajectory and surface modeling of the frame-level LSFs over the speech segments by polynomial functions is presented in Section III. A variable bit rate coder incorporating trajectory modeling of LSFs is presented in Section IV.

II. SEGMENTATION OF SPEECH

A segment is collection of fixed or variable number of frames. In case of variable length segments within each of which the speech parameters vary in a smooth manner, efficient segment based quantization methods can exploit most of the redundancy present in speech. In the present work, we consider variable length segmentation implemented as described below.

For segmentation, the input speech first is divided into voiced and unvoiced segments based on the frame voicing decisions obtained from the MBE analysis. Within each voiced and unvoiced section, the Euclidean distance between adjacent frames' LSF vectors is computed. Based on its value relative to a fixed threshold, a segment boundary is marked. This 2-step process is more effective in delineating homogenous regions compared with direct Euclidean distance based segmentation. In this segmentation method, the maximum segment duration may assume even large values. This may create problems for implementing the quantization of parameters. The maximum segment size is therefore limited based on the following specific considerations. In the vector quantization, the speech parameter of frames of segment are treated as vector elements and that vector is compared is with available codebook vectors and the best matched vector index is transmitted to the decoder. For each speech parameter and each input vector dimension, the corresponding codebook is required. If maximum segment size is large, large number of codebooks is required (large memory). Another important factor is the transmission of segment size, because at decoder for fetching a parameter vector using codebook index, proper codebook should be selected. Considering these two, the initial upper limit of segment size is fixed as eight (3-bit).

The other important parameter in the segmentation process is the threshold value, which is empirically selected to trade off between the two desirable aspects: maximizing segment duration, and avoiding over segmentation. A good trade-off would involve considering all the three performance issues: bit rate achieved, speech quality and memory. Bits required for LSF quantization are a major share of the total bits required and thus impacts the bit rate of the coder. Experiments are carried out on the training speech database (details provided at the end of paper) to identify the best combination of maximum segment size and threshold, considering LSFs' bit rate and their speech quality (in spectral distortion (dB)). Of course, the LSFs bit rate depends on the quantization scheme. In the present work, we propose to model LSFs with trajectory functions before quantization is done. By such modeling, only a reduced set of LSF vectors need to be quantized resulting in the reduction of bit rate. In the present work, a 50 % reduction in bit rate is targeted with respect to frame-based quantization.

Each combination of maximum segment size and threshold are tested on the training database and segments are extracted. The segments are modeled with trajectory functions, and spectral distortion is measured. For bit rate calculations, it is assumed that each LSF vector is quantized to 12 bits. The bit rate and speech quality for all combinations are measured. In this, the performance of maximum segment sizes 6 and 8 are found to be close. Further, considering the memory constraint, maximum segment size is selected to be 6.

A spectral distortion of 1 dB is generally considered to be acceptable for LSFs quantization. At the segmentation stage, quantization is not done so the threshold at which spectral distortion is 0.5 to 0.6 dB with modeling is selected for maximum segment size 6. This gives a threshold of 0.2. With the above parameter settings, the division of segments in the training database is as follows: segments with 1 frame (32 %), 2 frames (14 %), 3 frames (9 %), 4 frames (8 %), 5 frames (6 %), and 6 frames (32 %). The modeling of the segment LSFs by polynomial functions is presented in following section.

III. MODLEING BY POLYNOMIAL FUNCTIONS

In this section, a discussion of modeling by univariate polynomial functions is presented followed by the extension to bivariate polynomial modeling.

A. Univariate polynomial functions

In modeling of LSFs by univariate polynomial functions, the temporal trajectory of each LSF over a given segment is modeled by a low order polynomial function [2][3]. A polynomial function of order P is given by:

$$f_P(t) = a_P t^P + a_{P-1} t^{P-1} + \ldots\ldots + a_1 t + a_0; \quad (1)$$

where "t" represents the frame index. To find the coefficients of the polynomial fit on a segment of N frames, N equations are obtained and they are solved using least-squares optimization to obtain coefficients $\{a_P, a_{P-1}, \ldots., a_1, a_0\}$. For the given M LSFs, M unique polynomial functions will be constructed, one representing each LSF trajectory. In order to achieve efficient coding, it is necessary to encode the trajectory in some way. While the polynomial coefficients can be quantized, past experience with such an exercise is very limited in contrast to the vast research results on quantization of LSFs. Therefore, quantization is carried out on LSFs obtained by sampling the trajectories at P points rather than on the polynomial coefficients directly. On sampling these curves at P+1 points, P+1 feature vectors are derived. These feature vectors are encoded and transmitted to the decoder instead of the original N LSF vectors. Compression or gain in bit rate is obtained if P+1<N. The gain in bit rate depends on the P and N values.

For the given segment with N frames, depending on the compression required, P value is selected. For the 10-dimensional LSF vectors, 10 polynomial curves with P order will be formed. From each polynomial curve which approximates N frames LSF value, P+1 point on the curve are taken. This forms the P+1 feature vectors. These P+1 feature vectors are encoded and transmitted to the decoder. At the decoder, the P+1 feature vectors are extracted. The polynomial curves of order P are constructed using these P+1 vector. The curves are sampled at all original points and thus all original LSFs are obtained.

In the present implementation, the segment size ranges from 1 to 6. The segments with single frame are directly quantized (with 4096 codebook for the 10-dim LSF vector) without any use of polynomial functions. For other segments with frames two to six, polynomial functions are implemented. The polynomial order P is chosen based on segment size, targeting close to 50 % compression. In general, for even segment size P+1 is half of segment size and for odd segment size P+1 is floor (half of segment size+1). On obtaining the polynomial order for the segment, the polynomial curve with P order is constructed and the curves are sampled at P+1 points to extract P+1 feature vectors and each one is quantized with codebook of 4096 (12-bit). The indices are transmitted to the decoder where feature LSF vectors are decoded. The polynomial curves of P order are constructed using P+1 feature LSF vectors and curves are sampled at number of points as segment size and thus all LSFs of the segment are extracted. Like this, all LSF vectors are obtained with quantizing few feature LSF vectors and so gain in bit rate.

One important point to consider is the P+1 sampling point positions. As said earlier, the polynomial curve of order P approximated over N frames is sampled at P+1 points. Theoretically, the sampling points can be any P+1 points. But results are not same when implemented. The problem is better explained with an example. Let a segment of six frames with frames numbered as {1, 2, 3, 4, 5, and 6} is present. Second order polynomial curves are constructed. The curve is to be sampled at any 3 points to get feature LSF vectors for quantization. Let the sampled points be {1, 2, 3}, and these feature vectors are quantized and indices are transmitted. At decoder after reconstructing the polynomial curve, the curve is sampled at all six points. It is then observed that at points {4, 5, 6}, some spurious values occur. This is due to quantization step, if that step is not present, any three points will give appropriate values at all six points. To avoid this problem, the sampling points should contain the start and end points of the segment. The number of sampling points depends on the order P and it is best to choose them to be equally spaced as far as possible. For the case with segment size 6, one such sampling set can be {1, 3, and 6}. This method is followed for selecting the sampling points of all segment sizes. For other segment sizes the points chosen are: segment size 2 is {1}, segment size 3 {1, 3}, segment size 4 {1, 4}, segment size {1, 3, and 5}. The LSF vectors obtained for these segments are quantized with codebook of size 4096 (12 bits).

Table I
PERFORMANCE EVALUATION

Parameter	Results
Average bit rate (bps)	340
ASD (dB)	1.1
PESQ score	3.82
Outliers > 1dB (%)	53.65
Outliers > 2 dB (%)	2.53
Outliers > 5 dB (%)	0.00

The performance evaluation with these univariate polynomial functions is given in Table I. For the performance evaluation PESQ MOS [5] scores w.r.t LP modeled version and ASD (dB) are taken as measures. For comparison purposes, the method followed by Dusan [2] is also implemented. In this, a fixed segment size of 10 frames is modeled with fourth order polynomial functions. The ASD came as 1.7 dB at bit rate of 300 bps and the percentage of outliers > 5 dB is 3.7 % (quite high). It can be observed that the by implementing proposed segmentation method speech quality can be improved.

B. Bivariate polynomial functions

The univariate polynomial modeling described above captures the temporal correlations of each LSF independently of the others. If we can capture the joint time-frequency characteristics of LSFs and represent these with few coefficients or values then possibly a further reduction in bit rate can be achieved. Bivariate polynomial functions can model characteristics of LSFs along both time and LSF index axes. A surface is formed from the LSF values across both time and LSF index, and the surface coefficients are extracted and transmitted to the decoder. At the decoder, the surface is reconstructed by coefficients and the LSF information is recovered from the surface. The basic details of surface construction and the extraction of surface coefficients are explained below.

By using the bivariate polynomial function, a surface can be defined by a function f (x, y) for the given points $((x_1,y_1),z_1)$, $((x_2,y_2),z_2)$, $((x_3,y_3),z_3)$, ….., $((x_n,y_n),z_n)$. The variables x and y represent co-ordinates (frame index and LSF index) and z indicates the value (LSF value) at that point. The surfaces can be formed by the second or third order polynomial functions (further higher orders are also possible). The second order bivariate polynomial function is given as:

$$f(x, y) = c_0 + c_1x + c_2y + c_3xy + c_4x^2 + c_5y^2; \qquad (2)$$

In Equation (2), $\{c_0, c_1, c_2, c_3, c_4, c_5$ and $c_6\}$ are surface or bivariate polynomial coefficients to be estimated by least-squares fitting of the given data points.

The surface modeling is implemented selectively on homogenous segments (where sound variation is smooth) with 3 to 7 frames typically. The homogeneous segments are obtained by manual marking using the spectrogram of speech and covers about 70% of the frames. For segments with one to two frames, the input data requirement of minimum dimensions (3 x 3) matrix is not possible.

The surface modeling functions are applied in different ways. In the first case, using second order bivariate polynomial function a single surface is formed for entire segment and the surface coefficients are used at decoder to reconstruct surface and it is sampled to get all original LSFs. For this case, the results are very poor, because the surface coefficients (6 for 2nd order) are not able represent LSF information of 3 to 4 frames (30 to 40 elements for 10th order LP modeling) adequately. The ASD is observed to be above 1.6 dB and PESQ score is around 2.5 (maximum 4.5) in all cases, which are too high. The detailed results of one speech file are like this, total number of speech frames (433), number of segments marked (86), total number of frames in segments (287, 66 % of total frames), for second order bivariate polynomial functions, coefficients required are (86 x 6 =516). The average spectral distortion (ASD) obtained with this modeling is 1.9 dB and PESQ score is 2.3. Here, these ASD values are obtained just for modeling, if coefficients are also quantized, there will be further reduction in speech quality. The ASD value around 1 dB is acceptable value if quantization is also done. For modeling alone, ASD value around 0.5 dB can be treated as an acceptable value.

The problem just described is due to the attempt to fit a single smooth surface to a large region. It can be addressed by dividing the LSFs of the segment into two splits, the lower split contains the first five LSFs and upper split contains the other five LSFs. Each split is modeled with a second order bivariate polynomial function surface. Thus the LSFs information are represented with more surface coefficients leading to better reconstruction of the LSFs at the decoder. This gives better results than the previous case but still not good enough. For the same speech file considered earlier, the coefficients required are (86 x 12 = 1032) for second order surface modeling. The ASD observed is 1.0 dB and PESQ score is 3.4.

In the third case, we carry out further experimentation by increasing the number of surfaces further. Now, the LSFs of the segment are divided into three splits, the first one with three LSFs, and the second one with the next three and last one with the remaining four LSFs. Each split is modeled with a second order bivariate polynomial function surface. It is observed that the results are considerably improved over the earlier two cases. In this case for the same speech file mentioned in case 1, the coefficients required are (86 x 18 = 1548), ASD is 0.5 dB and PESQ score is 4.0.

One thing to observe is that from first case to third case, the number of coefficients is increasing in the order of {x, 2x, 3x}. If surface coefficients are going to be quantized then from first case to third case the number of bits required also increases as coefficients increased. The results are summarized in Table II, comparing the coefficients required: ASD (dB) and PESQ score for three cases mentioned.

To complete the full LSF quantization in this approach, the problem is quantization of surface coefficients. The surface coefficients are a new and untried set of parameters and their quantization properties are unknown. Thus quantization has not been incorporated in the present work but some suggestions are provided next.

Table II
COMPARISON OF DIFFERENT CASES OF BIVARIATE POLYNOMIAL MODELING

Number of surfaces per segment	**Coefficients required**	**ASD (dB)**	**PESQ score (wrt before modeling)**
One	516	1.9	2.3
Two	1032	1.0	3.4
Three	1548	0.5	4.0
Ten (univariate case)	1400	0.3	4.3

The indirect quantization of surface coefficients is possible via LSF quantization as follows. Consider the first case, where the whole segment is modeled by single second order bivariate surface. Let the segment have 4 frames (40 LSF values for 10th LP order). Assuming the use of second order bivariate polynomial functions, the six surface coefficients obtained should satisfy each LSF (z) value by its time (x) and LSF index (y) points; it should satisfy Equation (1). As mentioned, the coefficients obtained cannot be quantized directly. If any (3 x 3) matrix LSF points of the segment are transmitted to decoder, then at decoder the surface coefficients from the (3 x 3) matrix and so other LSF points can be obtained. The methods for the transmission of LSF points (quantization of LSFs) are well known. By this way, quantization process using bivariate polynomial functions can be feasible. The point to observe is, in the quantization of (3 x 3) matrix LSF points, the LSF values will more or less get modified, and then the surface coefficients (obtained by quantized (3 x3) matrix LSF points) and so other LSFs will also get modified. If the LSFs recovered at decoder are in acceptable range, then bivariate polynomial functions can be used for LSF quantization.

IV. VARIABLE BIT RATE CODER

In this section, we describe a fully quantized coder that uses univariate polynomial modeling of LSF trajectories as presented in Section III A. The parameters from MBE analysis are voicing decisions, LSFs, gain and pitch. The LSFs quantization is done by univariate polynomial functions as explained in Section III A. The quantization methods of the other parameters viz. voicing, gain and pitch are briefly discussed below.

Quantization of voicing decisions

From the MBE analysis, 12 band voicing decisions are obtained for each frame. The band voicing converted into 3 bit frame voicing index. To exploit the inter-frame redundancy, a study is conducted to observe the frequency of occurrence of particular voicing index patterns (consecutive frame voicing indices of segment). In the study, a few patterns are found to occur more frequently compared to other patterns. So, if these patterns are made into codebook, it will be enough for encoding frame voicing indices of input segment. This study is conducted for all segment sizes and codebooks are designed. These codebooks are used in encoding and decoding of voicing information of input segments. For selection of proper codebook at the decoder, segment size is also transmitted. This is the overhead present in this method.

Gain quantization

The gain values are highly correlated across the frames of the segment. For this, vector quantization is implemented. Log gain values are taken to reduce the dynamic range. Codebooks are generated for all segment sizes and gain quantization is implemented.

Pitch quantization

Vector quantization (VQ) is used for pitch parameter. Log pitch is taken to reduce dynamic range. In implementing VQ, it is observed pitch percentage error due to segment size 6 is more for smaller codebooks (less than or equal to 1024). To reduce this error, for the segment size six only, average pitch of segment and normalized vector obtained by subtracting the average pitch from the pitch values of the segment are quantized. This reduces the error and speech quality is also improved.

Postfiltering

Postfiltering is applied to reduce the coding noise introduced during the quantization process based on the fact that noise in the spectral valleys is more audible than that near the peaks. A pseudo-cepstrum based postfilter has been found to perform well and is incorporated in the decoder [4].

The overall average bit rate of the variable rate coder is given in Table III. The average PESQ MOS score is measured to be 2.7.

V. Discussion

Temporal trajectory modeling of LSFs across automatically selected homogenous segments of speech has been studied in the context of the MBE speech coder. As an extension to temporal trajectory modeling, we present a method to capture the joint spectro-temporal correlation via bivariate surface modeling of the LSFs. The issues that arise in the modeling and further quantization are discussed. Since full quantization has not yet been implemented in the surface modeling approach, any meaningful comparison with univariate trajectory modeling is possible only at the output of the polynomial modeling stage. Table II shows a comparison of the two approaches on the same speech data. We observe that the univariate modeling actually surpasses slightly the best performing bivariate surface modeled method, both in terms of speech quality as well as compression efficiency. This indicates that more research on the precise realization of the bivariate modeling concept is in order. Further investigations on the specific speech segments that may benefit more from one or the other type of segment modeling would be useful.

Finally, a complete variable rate coder based on univariate modeling of the LSF trajectories is presented. The average bit rate of the coder is 700 bps and it obtains an average PESQ MOS score of 2.7.

Appendix: Training data base for LSF quantization

The database comprises of 418 English sentences spoken by 49 speakers (both male and female). The speech of native English speakers was taken directly from TIMIT database and the rest were recorded in our lab. This database contains a total of 89702 frames of 20 ms duration.

Table III
BIT ALLOCATION TABLE

SR.NO	PARAMETER	AVERAGE BIT RATE(BPS)
1	Voicing	103
2	LSFs	340
3	Gain	112
4	Pitch	100
5	Segment size	45
	Total	700

References

[1] D. W. Griffin and J. S. Lim, “Multi-band excitation vocoder,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223-1235, August 1988.

[2] S. Dusan, J. Flanagan, A. Karve & M. Balaraman, “Speech coding using trajectory compression and multiple sensors,” *Proc. Int. Conf. on Speech & Language Proc*., Jeju, Korea, 2004.

[3] L.Girin, Long- term quantization of LSF parameters, *Proc. IEEE ICASSP 2007*, Honolulu, Hawaii, USA, 2007.

[4] R.S. Kumar, N.Tamrakar and P.Rao, “Segment based MBE speech coding at 1000 bps,” in Proc. *National conference on communication*, pp. 446-450, February 2008.

[5] ITU-T, “Rec P.862, Perceptual evaluation of speech quality (PESQ) an objective assessment of narrowband networks and speech codecs,” ITU, 2002.