COMPRESSION OF PROSODY FOR SPEECH MODIFICATION IN SYNTHESIS

Rashid Ansari *Wojciech Kurek*
Dept. of EECS (M/C 154), University of Illinois at Chicago,
851 S. Morgan St., Chicago, IL 60607, USA
E-mail: ansari@eecs.uic.edu,wkurek@eecs.uic.edu

ABSTRACT

In this paper, methods of compressing the prosodic information about a speaker's delivery – pitch, duration, and amplitude – are described. The objective is to use available or extracted knowledge of the spoken text along with the prosodic information to synthesize speech from a suitable inventory of stored basic units. Techniques for compressing pitch and amplitude of speech units using transform coding are investigated. Discrete cosine and sine transforms are found to be effective in compressing pitch and amplitude information respectively. In order to generate speech with these coded pitch and amplitude contours, the prosodic features of speech units stored in an inventory are modified using a method that was recently proposed to perform speech modification for concatenative synthesis. In this method, the stored speech unit is processed with a suitably shaped time-varying prefilter, whose parameters are chosen to have low sensitivity to pitch changes. The filtered signal is modified according to the required change in its prosodic structure, and then applied to the inverse of the prefilter. Examples of application of the proposed representation and modification of prosody to a variety of speech units are presented.

1. INTRODUCTION

There are several applications of speech synthesis in communication, such as in voice delivery of e-mail, voice response to database inquiries such as automated customer name and address information, mobile-environment communications, and speech compression [7, 9, 10, 20, 21, 24]. Our paper addresses issues in a non-traditional compression procedure based on synthesis. Very low-bit-rate speech compression can be obtained by exploiting the continuing improvement in the performance of speech recognition and synthesis systems. Speech recognition can be used to convert the information in the speech signal into the compact form of text. An auxiliary processor can be used to extract the prosodic information about the speaker's delivery – pitch, duration, and amplitude – from the speech signal. In some cases the corresponding text may be available in a database, and prosodic information may be provided to aid synthesis. Once the text and the prosodic information are available, one can synthesize speech from a suitable small inventory of stored basic units of speech as shown in Figure 1. Methods of performing the tasks of capturing the prosodic information of a speech utterance in a pitch-synchronous manner, representing it compactly, and using the compressed information to recreate the sound from an inventory of recordings at different pitch period are described here. Methods for coding prosodic information was earlier considered in [11]. In this paper we review the approaches described in [11], provide details of processing techniques used, and describe applications to multi-syllable utterances.

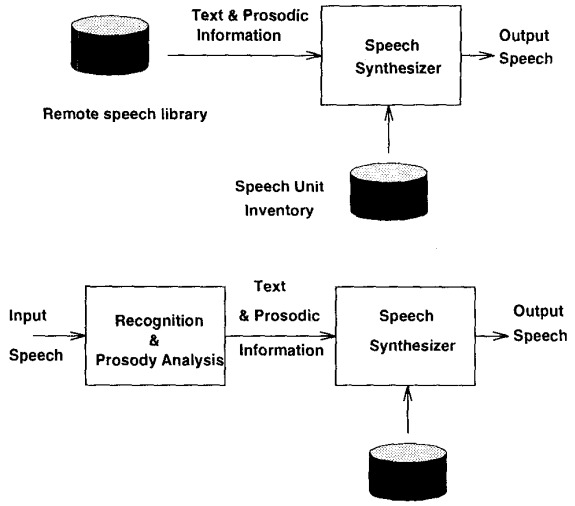

Figure 1. Application of stored prosody in speech communication

Speech coders based on source models are typically used for low bit-rate coding and produce intelligible speech at 2 kbps. These coder operations are based on a model of speech production, where parameters of the model are extracted from the signal being coded and the system is driven by a suitable excitation. If one disregards speaker characteristics and speech prosody, it has been noted that the fundamental information transmission rate for a human reading text is on the order of 100 bits/sec [23]. The approach considered here builds on this observation with the addition of prosodic information for speech synthesis in order to compress speech at significantly reduced bit rates. The approach relies on the use of concatenative synthesis in which consists of selecting a set of basic acoustic units, recording them in natural voice, and generating utterances

1058-6393/98 $10.00 © 1998 IEEE

by concatenating appropriately modified segments from the inventory of stored units. The key task of modifying stored speech is performed using the method described in [1] which was found to be effective in producing pitch modifications intended for generating clear and natural-sounding synthesized speech.

The overall compression method relies on effectively combining the techniques of speech recognition, synthesis and coding. Our effort represents the preliminary phase in addressing issues in selected aspects of the overall compression. The focus here is on issues related to the representation of the prosodic information, i.e. the pitch, amplitude, and duration of the speech units. Problems of imperfect recognition are not addressed. These include procedures of handling situations of recognition with low confidence, which may entail sending additional information resulting in increased bit rates. Perfect word recognition is not necessary in this application, as, for instance, there is no need to disambiguate homonyms, e.g "two" and "too".

To examine the effectiveness of the compression of prosodic information, we applied the procedure to compress the prosodic information of a number of stored recordings. These recordings include utterances of single words as well as strings of words. The prosodic information was extracted from these recordings and compressed using different fineness of quantization. Speech was then re-generated by applying the decompressed prosody to control the output signal using the inverse filter approach. The output was informally assessed to be satisfactory with prosodic information compressed at roughly 500 bps using transform methods. We also performed modifications of recordings of some single words by applying to it the prosodic information extracted from another recording of the same or different word but with a different prosodic structure. Details of the compressed representation and examples of speech modification are presented to explain the procedure.

2. PROCEDURE FOR SPEECH MODIFICATION

This section deals with the tasks to be performed by the decoder where specifications of the string of units to be concatenated and the prosody are available. It is assumed that the information about the speech units to be concatenated and the associated prosodic specification is available for speech generation at the decoder. Any suitable set of speech units can be used, though we have assumed that the inventory of speech units consists of demisyllables.

The speech modification used here is based on a new method [1] for altering the pitch, amplitude and duration of recorded female speech. The change in the pitch is the most critical task, and the discussion will be mainly centered on this task. The method in [1] was developed to overcome limitations in an otherwise promising technique called Residual-Excited Linear Prediction (RELP) [12]. In the new method, the stored speech unit is processed with a suitably shaped time-varying prefilter. The filtered signal is modified according to the required change in the fundamental frequency. The modified filtered signal is then applied to the inverse of the prefilter. Based on observations of spectra of multiple recordings of the same speech unit at different pitch frequencies, the magnitude response of the inverse filter was chosen to have a significantly less peaky structure than that which is typically obtained in linear predictive coding (LPC) [13]. Speech modifications using this method were found to be superior in quality to those obtained by RELP, while at the same time being less sensitive than RELP to errors in pitch marking.

2.1. Inverse filter approach

LPC, RELP, and PSOLA have been commonly used for modifying speech in concatenative speech synthesis [2, 3, 8, 9, 15, 16, 17, 21, 24, 25, 26]. Methods of speech modification have also been proposed in [5, 14, 18, 19].

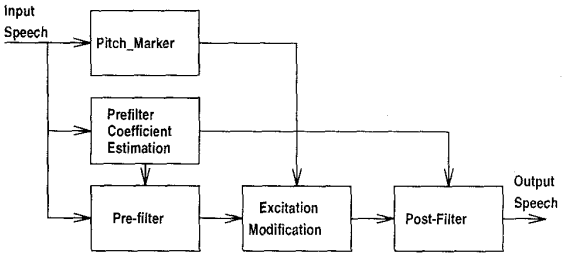

Figure 2. Inverse filter method for speech modification

In our work, we use a speech modification method which can be viewed as a generalization of the idea behind RELP (or more generally, LPC). It uses an inverse filter approach as shown in Figure 2. The filter parameters are not chosen to model the data by minimizing the residual energy, but to have reduced sensitivity to pitch modification. Speech recovery from the compressed data is based on the use of an inventory of speech units which are stored as i) system parameters per frame period, ii) a "residual" signal that is manipulated to generate the excitation signal, and iii) a file that provides the information about the frames. In the voiced portions of the stored units, the frames correspond to pitch periods.

The magnitude response is chosen to have a significantly less peaky structure than that which is typically obtained in LPC. The "residual" obtained with the new approach has significantly larger energy than that obtained with RELP and LPC. This causes larger discontinuities to appear in the excitation in the new method when the residual is modified with zero-padding to lower the pitch. In spite of this, the modified speech generated with the new method sounds better , as evidenced in the results of subjective tests.

In order to produce a less peaky filter response, the covariance matrix of the data in each frame is modified so that it produces an all-pole filter with a chosen lowpass response whenever the signal energy is reduced to zero. This approach, which can be interpreted as a form of regularization, is used in this speech modification method. The system parameters in the inverse filter approach are determined in a pitch-synchronous manner. A twelfth-order all-pole filter was used in the speech representation in each pitch period in the voiced portion. The intent is to provide reduced peakiness in the filter response to lower sensitivity to pitch alteration. To accomplish this, the covariance

matrix of the data in each frame was modified so that it produces coefficients of a chosen lowpass response whenever the signal energy is reduced to zero. This approach, which can be interpreted as a form of regularization, combined with bandwidth widening, produced significant improvement in performance over RELP. The covariance matrix is modified by adding a scaled set of modifier coefficients to the covariance matrix obtained from the data. The filter parameters are determined from the modified covariance matrix, $\mathbf{R}' = \mathbf{R} + \alpha\mathbf{C}$, where $\mathbf{R}$ is the covariance matrix obtained from the data, $\mathbf{C}$ is a symmetric Toeplitz matrix, and α is a scalar. A fixed $\mathbf{C}$ was used in the results reported here, with elements c_{1j} for $j = 1, \ldots, 9$ given in Table 1, and $c_{1j} = 0$ for $j > 9$. The remaining elements c_{ij} can be obtained from the symmetric Toeplitz structure of the matrix.

Table 1. Coefficients used to modify the covariance matrix

j	1	2	3	4	5
c_{1j}	0.2651	0.2010	0.1402	0.0686	0.0122
j	6	7	8	9	10
c_{1j}	-0.0165	-0.0207	-0.0129	-0.0045	0.0

2.2. Pitch, duration, and amplitude modification

Speech recovery requires specification of the string of concatenation units and the prosody. To produce speech with the specified concatenation units and prosody, the residual signals of the target units are retrieved from the inventory and altered suitably to create the excitation signals needed for the modified prosodic content.

The modification of stored parameters consists of two steps. In the first step, a new frame structure is created for the stored speech using the information available from the speaker's utterance. This typically involves inserting or deleting frames, depending on the nature of the desired and reference pitch and duration. In the second step, the size of each individual frame of the new residual signal (created in the first step) is changed to match the size of the corresponding frame of the speaker's new utterance. If a particular frame of the stored unit is longer than the corresponding frame of the input speech, then the residual signal of the stored unit is truncated to the size of the frame of the input speech. In the other case the residual signal is zero-padded to increase the pitch period to the target size. The LPC coefficient for the new frames are obtained from the inventory according to the frames created in the first step. The amplitude is changed according to a specified amplitude contour for the speech frames by adjusting the residual gain to yield the desired signal root mean-square (RMS) amplitude per frame. Finally the synthesized speech is created by using the data consisting of the new residual signal, the LPC coefficients, and the decoded RMS values, and applying them in the LPC synthesizer.

3. SPEECH PROSODY REPRESENTATION

In this section we consider the tasks to be performed in order to encode the prosodic information. The first task is to section the input speech into frames, where the frame sizes in voiced parts are equal to the pitch periods, while those in the unvoiced parts are equal. We used X-waves software to get a preliminary estimate of the pitch-marks, which was refined by processing with an adaptive narrowband filter and a median smoother. The RMS values of the signal per frame duration are also computed. To represent the pitch period information and the RMS values using a limited budget of bits, lossy compression need to be used. The high correlation in the data makes it suitable for transform coding [6, 7, 22].

The fact that discrete cosine transform (DCT) and discrete sine transform (DST) have excellent energy compaction for highly correlated data [6] makes them very suitable for coding the speaker's prosody parameters. In our work, the parameters needed for the prosodic representation are the length of the pitch period in the voiced section and the energy per frame duration which is given by the root mean square (RMS) value of the signal amplitude over a frame duration. Based on our investigation, DCT was found to be suitable for coding the pitch periods, DST was appropriate for representing the RMS values. The procedures for the coding described below.

3.1. Compression of pitch information

The voiced sections of the speech are identified and the pitch periods are estimated. Let $p[n]$, $n = 0, 1, \ldots, N-1$ denote the N pitch periods in a given voiced section of the speech. There is high correlation in the data as pitch periods vary slowly over a syllable duration. DCT was found to be effective in coding the data.

The DCT $P[k]$ of the sequence $\{p[n],\ 0 \le n \le N-1\}$ is defined as

$$P[k] = \alpha[k] \sum_{n=0}^{N-1} p[n] \cos \frac{\pi(2n+1)k}{2N}, \quad 0 \le k \le N-1 \quad (1)$$

where

$$\alpha[k] = \begin{cases} \sqrt{\frac{1}{N}} & k = 0, \\ \sqrt{\frac{2}{N}} & 1 \le k \le N-1. \end{cases} \quad (2)$$

The encoding procedure is shown in Figure 3. The DCT coefficients are quantized by dividing them by a scalar quantization factor q followed by rounding, and this yields the quantized coefficients $P_q[k]$, $0 \le k \le N-1$. It should be noted that one can also use a quantization vector, where a table of different quantization factors are used to provide a variable quantization step for the different DCT coefficients.

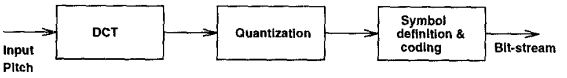

Figure 3. Procedure for coding pitch periods.

After quantization, the DCT coefficients are coded with variable-length coding using a procedure similar to that in the JPEG standard [22] for lossy image compression. In this method the run of zero coefficients is used to define events for coding. Also the possible values that quantized coefficients can take on are partitioned into power of 2 intervals (categories). The category numbers are then Huffman coded. Two different codes are used for encoding DC

and AC transform coefficients. The DC coefficient is coded based only on its value, whereas the code for non-zero AC coefficient is defined by its value together with the number of preceding zeros. Two additional special codes are used. One is used to indicate the end of data (EOD) and is generated after the last non zero value of data. The other (ZRL code) is used when the number of consecutive zero values exceeds 15.

3.2. Compression of RMS (amplitude) values

The RMS amplitude values vary significantly over the voiced portions of the speech, and not significantly over unvoiced regions. Therefore it is advantageous to code the RMS values separately for the voiced and unvoiced sections, where relatively low bit rate is needed for coding the unvoiced section.

Let $r[n],\ n = 0, 1, \ldots, N-1$ denote the N RMS amplitude values in a given voiced section of the speech. In our work multi-syllable utterances are sectioned at points in unvoiced parts. So the units to be coded are often single syllable voiced sections located between unvoiced sections. In this case the RMS data has small values at either end of the voiced speech. This makes it suitable for Discrete Sine Transform (DST) coding. The encoding procedure is similar to that shown in Figure 3, except that DST is used in the transformation instead of DCT.

The DST $R[k]$ of the sequence $\{r[n],\ 0 \leq n \leq N-1\}$ is defined as

$$R[k] = \sqrt{\frac{2}{N+1}} \sum_{n=0}^{N-1} r[n] \sin \frac{\pi(k+1)(n+1)}{N+1}, \quad 0 \leq k \leq N-1 \tag{3}$$

The coding procedure here is similar to that for coding pitch periods.

3.3. Estimate of bit rate requirements

An approximate estimate of the bit rate requirements in the proposed method is discussed. It should be noted that the rate will strongly depend on the quantization of the prosodic information. In representing the speech information, the text information corresponding to the recognition output will be estimated assuming five syllables uttered per second. If an inventory of two thousand demisyllable units is used then one needs roughly 11 bits/unit. Clearly with the use of variable-length coding this requirement can be reduced. The bit rate needed to represent the text information is roughly 110 bits per second (bps). The coded representation of the prosodic information for medium quantization was computed on a syllable basis. With medium quantization the average bit rate for pitch and RMS amplitude for a variety of vowel sounds and consonants is conservatively estimated to be 160 bits/syllable for medium quantization, and 80 bits/syllable for coarse quantization. This was based on results of representing a variety of sounds. Adding about 8 bits for syllable duration specification, a total of about 950 bps is needed for both text and prosody using medium quantization and about 550 bps using coarse quantization.

4. EXAMPLES OF PROSODY REPRESENTATION

Some examples of representation of the speech prosody are described in this section. The judgment of quality and difference in sounds is based on informal comparisons at this stage. An observation made regarding the quantization of prosodic features is that for a variety of sounds, the resynthesized speech produced from the quantized information sounds indistinguishable from the original for medium quantization of parameters.

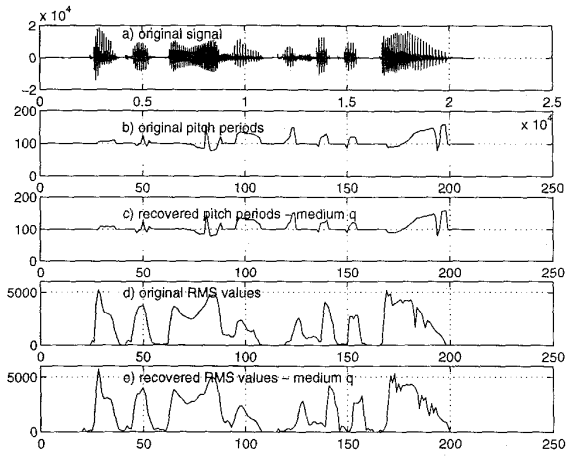

Figure 4. a) Original signal "Thank you for using Pacific Bell" b) original pitch periods c) pitch periods recovered after compression with medium quantization d) original RMS values e) RMS values recovered after compression with medium quantization

First we consider an example of extraction of the prosody of a sentence "Thank you for using Pacific Bell" in male voice. The voiced and unvoiced parts of the signal are identified and the voiced parts are pitch-synchronously divided into frames. Fixed sizes of frames are used for the unvoiced parts. The prosodic information was compressed with medium and coarse quantizers. In this example the compression rate was about 60% of that estimated above for medium quantization for which the differences were not perceivable. The signal with its original and decompressed parameters is shown in Figure 4 When speech is synthesized using coded pitch and amplitude contours without any change in prosody except that due to quantization, then there is no noticeable distortion even with significant quantization of DCT coefficients.

Next we consider the modification of a word "cups" uttered in female voice. There are two versions of the word: cups_l in a low-pitch voice and cups_h in a high-pitch voice. The durations of the utterances and the RMS amplitude contours are also different. The problem considered here is the modification of the signal cups_l in order to have the pitch, amplitude, and duration of the other utterance cups_h. The waveforms are shown in Figure 5. Another example is shown where the word cups_l is synthesized according to the prosodic specification of the signal for the

word "smith".

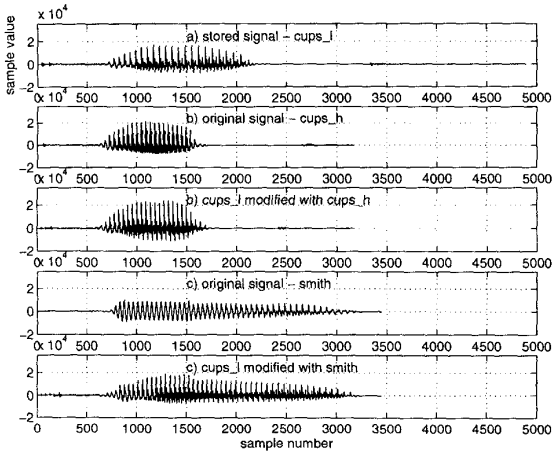

Figure 5. a) Stored low-pitch signal cups_l b) Original low-pitch signal cups_h, c) cups_l modified with prosody of cups_h, d) Original signal smith e) cups_l modified with prosody of smith

5. DISCUSSION

Methods of representing and modifying the prosodic features of speech are examined in this paper. These methods are intended for use in speech synthesis, especially for potential application in low-bit rate compression in conjunction with speech recognition. Transform coding is found to be effective in representing pitch and amplitude information. Other techniques of representing and compressing prosody information being investigated include a parametric representation of possible contours. This work represents a preliminary effort in handling this multi-faceted problem. Several issues need to be addressed. For instance, here the recognition is assumed perfect. Exceptions can be built for cases of recognition with low confidence, in which case traditional compression techniques can be intermittently used.

Acknowledgement

This work was supported in part by NSF under the grant IRI-9618887.

REFERENCES

[1] R. Ansari, "Inverse Filter Approach to Pitch Modification: Application to Concatenative Synthesis of Female Speech", *Proc. IEEE International Conf. Acoust., Speech, Signal Processing,* Munich, Germany, to appear, April 1997.

[2] G. Bailly and C. Benoit (Eds.), *Talking Machines, Theories,Models and Designs*, Elsevier, 1992.

[3] Sadaoki Furui. *Digital Speech Processing, Synthesis, and Recognition.* Marcel Dekker, New York, 1989.

[4] S. Furui and M. M. Sondhi (Eds.), *Advances in speech signal processing,* Marcel Dekker, New York, 1991.

[5] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.*, 40(6), pp. 497-516, June 1992.

[6] Anil K. Jain, *Fundamentals of Digital Image Processing,* Prentice Hall, Englewood Cliffs, NJ, 1989.

[7] N. S. Jayant and P. Noll, *Digital Coding of Waveforms,* Prentice Hall, Englewood Cliffs, NJ, 1984.

[8] D. H. Klatt, "Review of text-to-speech conversion for English," *J. ASA*, 82(3), pp. 737-793, Sept. 1987.

[9] W.B. Kleijn and K.K. Paliwal (Eds.), *Speech Coding and Synthesis*, Elsevier, Amsterdam, 1995.

[10] A. M. Kondoz, *Digital Speech - Coding for Low Bit Rate Communication Systems*, John Wiley & Sons, 1994.

[11] W. Kurek, R. Ansari, "Speech Prosody Representation for Synthesis and Compression," *Proceedings of Second International Conference on Multimedia Information Systems,* Chicago, IL, pp. 220-225, April 1997.

[12] M. J. Macchi, M. J. Altom, D. Kahn, S. Singhal, M. F. Spiegel, "Intelligibility as a function of speech coding method for template-based speech synthesis," *Proc. Eurospeech*, Berlin, Volume 1, p. 893-896, Sep 1993.

[13] J. D. Markel and R. M. Gray, *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[14] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-34(4), pp. 744-754, Aug. 1986.

[15] E. Moulines and F. Charpentier. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, 9, pp. 453-468, 1990.

[16] J. P. Olive and M. Y. Liberman, "Text-to-speech – An overview," *J. Acoust. Soc. Am.*, 78, S6, 1985.

[17] O'Shaughnessy, D. (1987), *Speech Communication: Human and Machine*, Addison Wesley, New York, 1987.

[18] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-30, pp. 374-390, June 1981.

[19] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation,"*IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-34(6), pp. 1449-1464, Dec. 1986.

[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, NJ, 1978.

[21] D. B. Roe and J. G. Wilpon (Eds), *Voice communication between humans and machines*, National Academy of Sciences, 1994.

[22] K. Sayood, *Introduction to Data Compression*, Morgan Kufmann Publishers, 1996.

[23] R. W. Schafer, "Scientific Bases of Human-Machine Communication by Voice", *Voice communication between humans and machines*, National Academy of Sciences, 1994.

[24] A. Syrdal, R. Bennett, and S. Greenspan, *Applied Speech Technology*, CRC Press, Boca Raton, FL, 1995.

[25] J. Van Santen, R. Sproat, J. Olive, and J. Hirshberg, editors. *Progress in Speech Synthesis*, Springer Verlag, New York, 1995.

[26] I. H. Witten. *Principles of Computer Speech*, London: Academic Press, Inc., 1982.