IMPROVING HIGH QUALITY TTS USING CIRCULAR LINEAR PREDICTION AND CONSTANT PITCH TRANSFORM

S. Shukla, T.P. Barnwell III

Georgia Institute of Technology,
School of Electrical and Computer Engineering
Atlanta, GA
{shukla,tom@ece.gatech.edu}

ABSTRACT

Current high quality concatenative TTS systems are based on unit selection from a database that is contextually and prosodically rich. These systems are computationally expensive and require a very large footprint. This paper presents a new method for representing speech segments that can improve the quality and scalability of concatenative TTS systems. The circular linear prediction model combined with the constant pitch transform provides a robust representation of speech signals that allows for limited prosodic movements without perceivable loss in quality. A method is presented for constraining the LSF tracks of speech segments to realize pitch modifications with minimal artifacts. The results of formal listening tests demonstrate that limited prosodic modifications can produce speech from fewer units whose quality equals or exceeds large database unit-selection systems. Additionally, this method is used to realize high quality emphasized speech.

Index Terms— Speech synthesis, speech communication, linear predictive coding, speech processing, speech intelligibility.

1. INTRODUCTION

Over the last several years, the area of concatenative text-to-speech (TTS) synthesis has seen significant advances in terms of voice quality and intelligibility. Commercial and research systems by AT&T, SVOX, Cepstral, Festival, the MBROLA Project, and others provide viable solutions for interactive applications that would otherwise require a real human voice. These systems can be classified under three general categories: diphone synthesis, unit selection synthesis, and limited-domain synthesis.

Diphone synthesis is based on the concatenation of recorded units at the midpoint of each phoneme. This has been a preferred method due to its ability to synthesize an unlimited vocabulary using a very small footprint (1000 to 2500 units). However, since generally only one instance of each unit is stored in the database, significant prosodic (segmental pitch and duration) modifications are required for intelligibility. The modifications are applied using a speech model such as RELP [7][9], or MBR-PSOLA [1] to parameterize the units. For existing speech models, prosody modifications introduce artifacts of varying degree depending on the extent of modification, class of unit (vowel, consonant, fricative, etc.), and speech model. This coupled with the large number of segment boundaries inherent in diphone synthesis, results in speech that sounds unnatural.

Recently, TTS based on unit-selection synthesis has gained wide acceptance due to its ability to produce "customer quality" speech [2]. The synthesis database, consisting of numerous instances of each unit, is rich in context, spectral characteristics and prosody. This reduces or even eliminates the necessity for prosody modifications and boundary smoothing. For achieving "customer quality" speech, however, a labeled corpus of over 10 hours of speech is required [3]. Though these systems are scalable, the quality is predictably compromised as the database is reduced. “Limited-domain” TTS is a popular version of unit-selection synthesis used in current applications such as telephone banking, information kiosks, etc. Since the vocabulary and subject are limited, the database consists of larger units of varying prosodic content to achieve "natural quality" with a relatively smaller footprint than unlimited unit-selection TTS.

Current implementations of unit-selection TTS typically use RELP, MBR-PSOLA, or the Harmonic plus Noise Model (HNM) (a variant of the Sinusoidal Model), for representing the speech and applying prosodic variations. Though perceptually high quality synthesis is achievable, these methods have inherent modeling errors. These errors can produce audible artifacts at segment boundaries when applying prosodic modifications. Circular Linear Prediction combined with the Constant Pitch Transform (CLP/CPT) provides a more robust model for representing speech that is, theoretically, free of modeling errors. This representation can enhance the performance of the current TTS systems by providing a method for high quality prosodic modification. Specifically for unit-selection and limited-domain TTS, this method can reduce the storage requirements by relaxing the degree of prosodic richness required in the segment database.

2. THE CLP MODEL AND THE CPT

The circular linear prediction model and the constant pitch transform have been detailed by [5], and presented for improving speech coding applications by Ertan and Barnwell [6]. CLP is a windowless approach to LP modeling, based on the key assumption that the analysis frame is exactly periodic. Performing the analysis on pitch-synchronous, non-overlapping frames with fractional pitch resolution can satisfy this assumption. As in traditional LP modeling, the coefficients of the all-pole filter, *A(z)*, are determined by minimizing the squared error of the modeling region, resulting in the linear equations:

$$\sum_{i=1}^{P} a_i r(i,j) = -r(0,j) \qquad j=1..p, \tag{1}$$

where *r(i,j)* is the covariance of the analysis frame, *s(n)*. However, since the CLP analysis frame can be viewed as an infinite periodic signal of period, T_0, the covariance, *r(i,j)*, simplifies to:

$$r(i,j) = r(k) = \sum_{n=0}^{T_0-1} s[n]s[((n+k))_{T_0}], \qquad (2)$$

where $k=|i-j|$ and $((.))_{T_0}$ is modulo T_0 operation. Since *r(k)* is an autocorrelation function, the CLP coefficients can be determined using the autocorrelation method without making any assumptions about the signal outside the modeling region. This provides the advantage of greater accuracy in modeling of pitch periods while maintaining the stability and efficiency of autocorrelation LP modeling. Note that for unvoiced speech, the pitch period does not exist and an arbitrary frame length can be used. For fractional pitch resolution, a pitch period of length T_0+f, where T_0 is the number of integer samples and *f* is the fraction, is upsampled by a factor, *N*, so that $N(T_0+f)$ is an integer.

The CPT is a method for modifying all of the pitch periods of a segment to a fixed length, T_C. This transform serves two key purposes: (1) convert the fractional resolution pitch periods of the CLP analysis residual signal to integer lengths, and (2) create a segment database of fixed pitch periods to facilitate pitch modifications using the inverse CPT during synthesis. As shown in figure 1, each pitch period of the residual signal, $e_K(n)$, is upsampled by the constant pitch, T_C, filtered to prevent aliasing, and downsampled by the original fractional period. A key consideration for T_C is that it should always be greater than the largest allowable analysis pitch period so as to prevent aliasing when downsampling. The very large computational complexity of this representation can be of concern for some applications. However for TTS, the analysis is conducted offline during the database preparation stage.

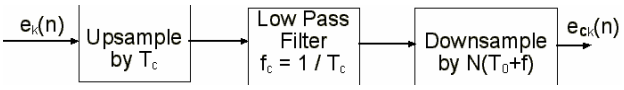

Figure 1: Block diagram of Constant Pitch Transform

3. TTS IMPLEMENTATION USING CLP/CPT

The implementation of TTS for this research was conducted using the phonetically labeled CMU Communicator speech database [4]. This unit-selection TTS database, designed for the limited-domain application of a travel reservation system, has been made available for research purposes. The phoneme boundaries for the segments in the database have been determined automatically, and were hand corrected to avoid synthesis errors. In addition, boundary locations for voiced and unvoiced speech labels were determined.

3.1. Analysis Phase

One of the key implementation problems with synthesizing artifact-free, "natural quality" speech is the variations in acoustical characteristics of the segments, caused by changes in the recording environment, when recorded over a long period of time. This results in audible spectral and dynamic variations in the concatenated units. This research implements an equalization method based on matching low order (4^{th}) LP spectra of all segments in the database to the spectra of a target segment.

3.1.1. Pitchmark Placement

The pitch-synchronous frame boundaries, or pitchmarks, are determined by, first, calculating the pitch track of each segment using a pitch detector. Pitch epoch locations are then automatically determined in the residual domain based on the method by Smits and Yegnanarayana [8]. The residual signal is determined using windowed autocorrelation LP analysis. It is well known that real speech signals are not perfectly periodic and even at the fractional resolution there will be slight errors in the pitch period. During prosodic modifications, these errors can be magnified, resulting in audible artifacts. To minimize the effect of these errors, the pitch cycle boundaries are set not at the "true" beginning of the pitch epoch, but rather at the low instantaneous energy region at the onset of the pitch epoch. This is determined by locating the first zero crossing prior to the true beginning of the pitch epoch. Since accurate pitchmark placement is vital to the success of CLP analysis, hand correction of placement errors is necessary. Finally, segment boundaries are truncated so that each database unit begins and ends exactly on a pitchmark. This ensures that units are concatenated only at pitchmark locations during synthesis.

For unvoiced and partially voiced signals, the pitch periods of adjacent voiced signals are interpolated to maintain a smooth pitch track. Though this is sufficient for unvoiced speech, it can introduce modeling errors in partially voiced speech. However, smooth pitch and duration tracks are critical for matching to target prosody parameters. Sudden changes in the pitch and duration factors can lead to more artifacts than due to the modeling errors associated with partially voiced signals.

3.1.2. Constant Pitch Segment Database

For fractional resolution of the pitch, each speech unit in the database is first upsampled by a factor, *N*, which is determined by the desired resolution. It has been determined that one decimal place of precision ($N=10$) provides sufficient accuracy for CLP modeling [5][6]. CLP analysis is performed on a number of fractional pitch periods within a predetermined range of the original pitch estimate. A range of ± 2 integer samples was used for this implementation. The pitch period length that maximizes the LP gain is chosen as the best fractional pitch. The resulting coefficients are used to calculate the residual signal by performing circular inverse filtering of each pitch period. Note that the LP coefficients, *A(z)*, need to be converted to $A(z^N)$, since the signal, *s[n]*, has been upsampled by *N*. Finally, the residual signal is transformed to a constant integer pitch by the CPT. The segment database consists of the CLP coefficients, the original pitch period length of each frame, and the constant-pitch residual signal. Additionally, phoneme boundaries within the segments are stored in terms of pitch period indices.

3.2. Synthesis Phase

3.2.1. Prosody Matching

The target prosody for TTS is generally determined by the Natural Language Processing module in a TTS system using either a set of rules or statistical methods. For the purpose of this research the target prosody was extracted from real speech recordings of the utterances to be synthesized. The pitch modifications are performed prior to duration modifications, using the inverse CPT, because the resulting residual signal will also change in duration. The target pitch track for a given segment is mapped to every frame so that there is a target pitch T_{0i}' for every pitch period. The

inverse CPT is then implemented for each frame, *i*, *as* shown in figure 2.

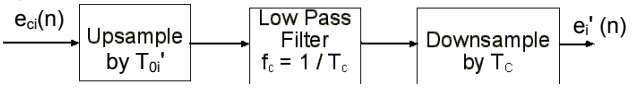

Figure 2: Pitch Modifications Using the Inverse CPT

For each segment residual signal of length, *N*, a new set of pitchmarks, M_i', is created, as follows:

$$M_i' = M_{i-1}' + T_{0i}' \qquad 1 < i < N/F_C, \tag{3}$$

where $M_0' = 1$ and F_C is in units of (*1/samples*). The target durations are realized by either repeating or deleting entire frames of the residual. The duration factor for each phoneme is calculated from the target duration and the current phoneme duration. Based on this duration factor, frames of the phoneme are either repeated or deleted to increase or decrease the duration, respectively. A mapping function, *MAP*, is derived that maps the original frame indices, *i*, to the indices for the duration modified frames, *j*. The mapping function is applied to the pitch periods (frames) of segment, T_{0i}', to create a new set of pitch periods, $T_{0MAP(j)}'$, and a third set of pitchmark locations, M_j'', are calculated as follows:

$$M_j'' = M_{j-1}' + T_{0MAP(j)}' \ , \tag{4}$$

where $M_0''=1$. The residual, *e'(n)*, is also modified to, *e''(n)*, in a similar manner by applying the mapping function to concatenate the repeated and deleted pitch periods resulting in *e''(n)* as in (5).

$$e_j''(n) = e_{MAP(j)}'(n) \cdot \tag{5}$$

Though the CLP/CPT speech representation allows for prosody modifications with minimal artifacts in the synthesized speech, the extent of the modifications is limited. For this reason, thresholds have been derived to limit the pitch-scale factor and duration factor. Based on informal subjective testing, different thresholds were derived for the various phoneme types. For example, vowels were limited to duration fluctuations of *±30%*, fricatives were limited to *±18%*, and stop consonants were limited *±5%*. Additionally, to prevent sudden inflections in pitch, the pitch-scale factors are smoothed by a 6-tap moving average filter.

3.2.1. Unit Concatenation and Synthesis

Due to the assumption of exact periodicity, unit concatenation can be achieved simply by placing the CLP synthesized segments end-to-end. The analysis method of the CLP/CPT representation ensures that pitch epochs are aligned at the boundaries of voiced and partially voiced speech segments. Hence, smooth junctures at the concatenation points can be realized with no interpolation of parameters or the residual signal. After the new pitchmarks, M_j'', and residual signal, *e''(n)*, representing the target pitch periods and durations, have been created, the segments are synthesized by the CLP synthesis methods described by [5].

3.3. Constraints on the LSF Tracks

Ansari [7] observed that the peakiness and poor bandwidth estimates inherent in LP spectra affects the quality of speech when the pitch is modified. This observation resulted in an improvement to RELP-based TTS where the LP model was modified to produce a less peaky magnitude response. As opposed to modifying the model itself, a method has been implemented for selectively widening the bandwidths of extremely narrow-bandwidth formants by constraining the line spectral pair (LSP) coefficient tracks so

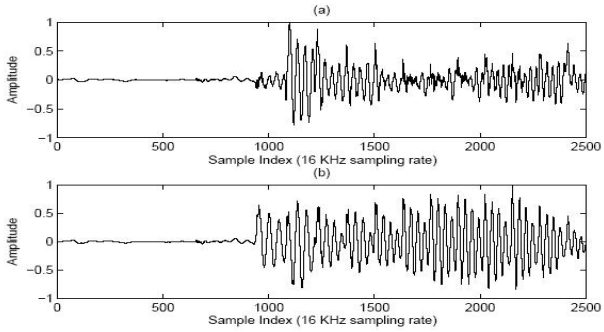

Figure 3: CLP/CPT synthesis of voicing transition with pitch modification (a) before and (b) after applying LSF constraints

that undesirable artifacts due to pitch modifications are reduced. Thresholds are applied to the LSP coefficients to maintain a minimum distance between each pair. Though, this minimum spectral distance threshold, F_m, is a constant value, it is not implemented in a strict sense. While maintaining a certain distance between each pair of coefficients, an acceptable distance between adjacent pairs (P_{i+1}, Q_i) must also be maintained. This procedure is implemented iteratively, first applying the threshold between each pair of LSP coefficients (P_i and Q_i), and then between adjacent pairs (P_i and Q_{i-1}). Before applying the threshold, the midpoints between each pair of coefficients, $C_{i,i}=(P_i+Q_i)/2$, and the midpoints between adjacent pairs, $C_{i,i+1}=(P_i+Q_{i+1})/2$, are determined. Then, the threshold is applied between each pair as shown in (6) and (7).

$$P_i = MAX(MIN(P_i, C_{i,i} - F_m), C_{i-1,i}) \cdot \tag{6}$$

$$Q_i = MIN(MAX(Q_i, C_{i,i} + F_m), C_{i,i+1}) \cdot \tag{7}$$

After applying the threshold to all the pairs, a second pass is made on the LSP coefficients to apply the threshold to the adjacent pairs as shown in (8) and (9).

$$Q_i = MAX(MIN(Q_i, C_{i,i+1} - F_m), C_{i,i}) \cdot \tag{8}$$

$$P_{i+1} = MIN(MAX(P_{i+1}, C_{i,i+1} + F_m), C_{i+1,i+1}) \cdot \tag{9}$$

Figure 3 above shows that the artifact caused by pitch modification at the unvoiced-voiced speech transition (a) is significantly reduced by applying the LSF constraints (b).

4. FORMAL LISTENING TESTS

Two subjective listening tests were conducted to determine the quality of utterances synthesized by the CLP/CPT method with prosodic modifications. In the first test, subjects compared utterances synthesized by this method to the same utterances produced by unit-selection synthesis with no prosodic modifications. The second test analyzed the ability for the CLP/CPT representation to synthesize speech with increased emphasis. In this test, the subjects selected a preference between unmodified speech and speech emphasized using CLP/CPT. The control and test utterances in both subjective tests were synthesized from the CMU Communicator limited-domain TTS database. The subject matter of the utterances was that of a simulated dialogue with a travel reservation system. To simulate a real use case (making travel reservations while driving a car), road noise at highway speeds was added to the reference and test utterances.

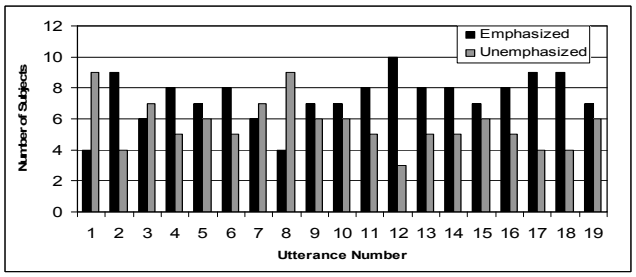

Figure 4: Results of Subjective Test of Emphasis Realization

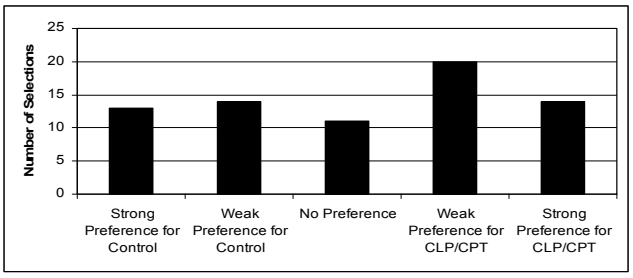

Figure 5: Results of Subjective Comparison of CLP/CPT and Unit Selection Synthesis

4.1. Test Method

For the first test, the target pitch and duration values for each phoneme of the test utterances were extracted from the control utterances. The synthesis database was created by selecting units from the CMU Communicator database. The units were selected such that each of the control utterances can be synthesized and only one instance of each unit exists. Since, these units are not the same as the ones used to create the control utterances, the prosodics will naturally differ. Using the CLP/CPT method, the database was analyzed and the test utterances were synthesized with the target pitch and durations. There were 12 subjects and 6 sets of control and test utterances resulting in a total of 72 responses. For each set of utterances, the subjects could select from 5 preference choices: strong or weak preference for the control utterance, a strong or weak preference for the CLP/CPT synthesized utterance, or no preference.

For the second test, the control utterances were unmodified utterances synthesized by the CMU Communicator and the test utterances were created by CLP/CPT resynthesis of the control utterances with the prosody of key words emphasized to increase intelligibility. There were 13 subjects and 19 pairs of utterances. This test was conducted as an A-B comparison test, in which the subjects selected a preference for or against emphasis. For both tests, road noise was added to all utterances after synthesis.

4.2. Results and Analysis

The results of the first test, given in figure 4, show a relatively even distribution, with a slight preference by the subjects for the utterances synthesized by the CLP/CPT method with prosodic modifications. To analyze the significance of this result, a one way analysis of variance (ANOVA) was performed with respect to the distribution of preferences. The results of the ANOVA revealed a 91% confidence interval indicating that the slight preference for CLP/CPT shown by the data is not highly significant.

Figure 5 gives the distribution of the results of the second test for each utterance pair. Overall, the CLP/CPT emphasized utterances were preferred 57% to 43% over the "natural" unemphasized utterances. In this case the T-test showed high statistical significance (>99.5% confidence interval) supporting the preference for emphasis using CLP/CPT. This result indicates that the prosody modifications for realizing emphasis do not noticeably degrade voice quality yet may improve intelligibility.

The amount of database reduction and improvement in quality that is achievable is difficult to quantify because it varies based on the contents of the database. For example, for limited domain unit-selection TTS applications that strive for "natural quality", the database often has numerous instances of the same segments with very slight changes in prosody. Such a database can be reduced significantly using the CLP/CPT representation to modify prosody.

5. CONCLUSIONS

Current unit-selection based TTS systems synthesize very high quality, artifact-free speech by using a large, prosodically rich speech database. Historically, applying prosodic variations with a speech model resulted in artifacts and unnatural speech quality. This research demonstrates that limited prosodic variations can be realized by the CLP/CPT method without noticeable degradation in speech quality. By applying this method to existing systems, the number of prosodically varied instances of each unit can be reduced to achieve similar synthesis quality. Alternatively, the CLP/CPT method can be utilized to further improve the prosodics and even increase emphasis of syllables.

6. REFERENCES

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, 1997.

[2] A. Hunt and A. Black, *"Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," IEEE ICASSP 96*, Atlanta, Vol.1. pp. 373-376, 1996.

[3] J. Schroeter, *"The Fundamentals of Text-to-Speech Synthesis," VoiceXML Forum*, Vol.1, Issue 3, March 2001.

[4] A. Rudnicky et al, *"Creating Natural Dialogs in the Carnegie Mellon Communicator System," Proceedings of Eurospeech*, Vol.4. pp. 1531-1534, 1999.

[5] S. Shukla, E. Ertan, and T.P. Barnwell, *"Circular LPC Modeling and Constant Pitch Transform for Accurate Speech Analysis and High Quality Speech Synthesis," IEEE ICASSP 02*, Vol.1. pp. I-269-I-272, 2002.

[6] A. Ertan and T. Barnwell III, *"Circular linear prediction modeling for speech coding applications," Proceedings of 37th Asilomar Conference on Signals, Systems and Computers*, pp. 290–294, 2003.

[7] R. Ansari, *"Inverse Filter Approach to Pitch Modification: Application to Concatenative Synthesis of Female Speech," IEEE ICASSP 97*, Vol.3. pp. 1623-1626, 1997.

[8] R. Smits and B. Yegnanarayana, *"Determination of Instants of Significant Excitation in Speech Using Group Delay Function," IEEE Trans. on Speech and Audio Processing*, Vol.3, No.5, pp. 325-333, 1995.

[9] M. Macon, A. Cronk, J. Wouters, A. Kain, *"OGIresLPC: Diphone synthesizer using residual-excited linear prediction". Tech Report CSE-97-007*, Dept of Comp Sci, Oregon Grad Institute of Science and Technology, Portland, Sept. 1997