Pitch Control in Diphone Synthesis

H. T. Bunnell, D. Yarrington and K. E. Barner

Applied Science and Engineering Laboratories
A. I. duPont Institute and The University of Delaware
Newark, DE 19716

Abstract— **A hybrid time domain and LPC approach to speech pitch control is developed. This approach uses a low order LPC analysis and residual excitation to alter pitch period length during voiced speech. This approach differs from standard residual excited LPC in that LP reconstruction is applied only during voiced segments. Listening tests were used to compare PSOLA and the hybrid method under conditions of increasing or decreasing F0 in natural speech tokens. The natural speech was recorded by two talkers, a female adult and female child. Results suggest that, while the overall performance of the two methods is similar, the methods differ in their effectiveness with direction of F0 shift and over talkers.**

Keywords— **Speech, diphone synthesis, pitch modification.**

I. Introduction

Control of intonation and timing is difficult in diphone concatenation synthesis, especially when one aim of the synthesis is to capture and present the voice characteristics of a specific talker. One method for controlling F0 and duration in diphone synthesis, the Pitch Synchronous Overlap Add (PSOLA) technique (e.g., Hamon, *et al.*, 1989) has been shown to produce high quality copy synthesis, and to provide control of both F0 and duration. PSOLA is implemented by manipulating windowed epochs of the time signal. In voiced speech, each epoch is derived by applying a Hanning window centered on the onset of a pitch period and having a length corresponding to twice that of the pitch period.

However, the PSOLA method is sensitive to pitch tracking errors and can introduce distortion when F0 is changed by too great a degree. Moreover, PSOLA has been found to produce greater distortion than residual excited LPC methods of F0 control, at least under some conditions (e.g., Macchi, *et al.*, 1993). Here, we derive a hybrid LPC and time-domain pitch control method that uses residual excited LPC reconstruction for voiced segments only. In diphone synthesis applications, Voiceless segments are processed as per the PSOLA method.

The authors wish to thank Kathy Samworth for her assistance with data analysis, and Joshua Gallant for his work on the EGG based pitch tracker. This work was supported by NIDRR and by the Nemours Foundation.

II. Hybrid LPC Method

The present pitch modification methods derive from efforts in our laboratory to develop general prosodic modification capabilities for use in diphone speech synthesis and for enhancing dysarthric speech. The processing method, like PSOLA, must therefore provide independent control of F0 and segment duration. While F0 modification applies only to voiced segments, more general duration control applies to all segments, whether voiced or not. The present method is a hybrid in the sense that it processes voiceless segments (for duration control) using simple time–domain processing similar to the PSOLA method, but uses residual-excited LPC (RELP) processing to reconstruct pitch periods for voiced segment processing. The present report is concerned primarily with the performance of the RELP component of the hybrid LPC method (hereafter HYBRID) in controlling pitch.

A. Algorithm Description

To alter the pitch in the voiced speech case, an LPC analysis and reconstruction method is used, Figure 1. The speech is first partitioned into frames consististing of a single pitch period starting at a zero crossing associated with glottal closure appended to a brief (2.0 msec. in the current study) "overlap" segment of prior signal context. The overlap segment is used during signal reconstruction for inter-frame smoothing to reduce transients. A low order ($8'th$ order filter and 20.050 kHz speech in the present study) LPC model is fitted to the Hanning windowed, first differenced, pitch frame (including overlap interval) using the autocorrelation method. The residual error signal is then computed from the unwindowed pitch frame. The low-order LPC model extracts just the most prominent spectral characteristics of each frame, leaving

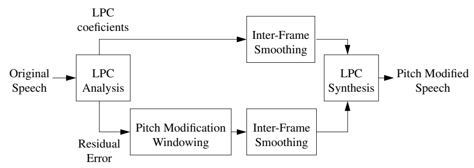

Fig. 1. Block diagram of HYBRID-LPC pitch modifier for the case of voiced speech.

the excitation (presumably) and fine spectral detail of each frame to be represented in the residual. Regardless of the model order, the initial part of each pitch period is reconstructed exactly from the RELP synthesis.

F0 modification is achieved by exciting the LPC model with a duration adjusted residual. The duration of the residual is adjusted by windowing, and for the case of decreased pitch, zero padding. The window applied to the residual is defined by:

$$w(i) = \begin{cases} 1.0 & \text{for } 0 < i < 0.75T \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(i - 0.75T)}{0.25T}\right)\right) & \text{for } 0.75T \leq i < T \end{cases}$$

where T is the desired output pitch period duration, or the full duration of the original (input) pitch period, whichever is smaller.

For increased F0 (T less than original pitch period duration), the windowed (and shortened) residual is used to excite the synthesis filter. To decrease F0, the windowed residual error is zero padded to the desired duration and then used to excite the synthesis filter.

The pitch in voiced speech is thus modified by altering the duration of the LPC model residual error and exciting the model with the altered residual. Additionally, the LPC parameters and residual are inter-frame smoothed through linear interpolation during the overlap interval prior to the nominal pitch period onset. This smoothing functions to decrease ringing and other artifacts at pitch period boundaries.

III. Listening Test

Both the HYBRID method described here and the PSOLA method have been implemented in our laboratory as part of an experimental diphone speech synthesizer for multiple voices, including particularly, women and children. Informal listening comparisons of the two methods for pitch control led us to the subjective impression that the HYBRID method produced slightly, but noticeably, less distortion, especially when dealing with high-pitched speech. Since this subjective impression was consistent with reports that RELP methods outperformed PSOLA under at least some conditions, we sought to verify the impression with formal listening tests designed to compare the segmental intelligibility of speech processed by the two methods.

A. Method

A. *Stimuli*

To avoid any possibility of artifacts related to the diphone synthesizer itself, the present study was carried out using the HYBRID and PSOLA methods for pitch modification on digitized monosyllabic words recorded by talkers for this experiment. Forty-nine non-words were selected from the list used by Spiegel, et al. (1990) for assessing consonant intelligibility. The words chosen were those which included the English voiced and voiceless stops, singly or in clusters, in either word onsets or offsets/affixes. The 49 words were recorded by two female talkers, one an adult and one a 10 year old child. With each talker seated in an IAC booth, two channels of data were recorded to DAT tape: the audio signal; and the output of a Glottal Enterprises electroglotagraph (EGG). The EGG output was used to assign pitch period markers for voiced regions of the speech. For processing, the audio output of the DAT recording was filtered and sampled at 20.050 kHz with 16-bit resolution on an SGI Indy computer. Individual tokens were located, excised from the recording, and saved for subsequent processing. Three words were found to contain dysfluencies or were simply mispronounced by one or the other of the two talkers. One of these words was removed from the corpus of one talker, but the other words remained in the corpus for the listening sessions. None of the three words for either talker were included in the final data analysis.

Four F0-shifted stimuli were generated from each natural token by shifting the F0 of the original up by a factor of 1.6, or down by a factor of .7 using each of the two processing methods (PSOLA and HYBRID). In all there were thus 485 stimuli (2 TALKERS by 5 F0-SHIFT conditions (original speech plus two shift directions and two shift methods) by 49 WORDS for one talker and 48 for the other).

B. *Procedure*

Twenty volunteer listeners heard each of the stimuli over headphones in an IAC booth. Each word was presented once, after which the listener had as much time as necessary to enter a response on a keyboard. All subjects were told before the test began that they would be hearing non-words and that they should try to spell each nonsense word "phonetically." Stimulus presentation order was randomized for each listener and the experiment was broken into three roughly equal length blocks of trials with a rest break between each trial block. Most subjects completed the listening task within 90 minutes.

C. Data Reduction and Analysis

We compiled a table of all unique spellings given in response to each of the 46 words that were produced correctly by both talkers. Each unique response was assigned a classification which indicated whether (a) the entire response was a phonetic match to the stimulus item (word-correct), (b) the onset consonant or cluster of the response matched that of the stimulus item (onset-correct), or (c) the offset consonant or cluster of the response matched that of the stimulus item (offset-correct). Three judges examined the classification items in the table to arrive at a consensus. For the present paper, only word-correct scores will be considered (onset and offset data were found to provide similar, although weaker, results).

The dependent measure for the repeated measures analysis of variance was the difference in percentage of words (N=46) correct between the unaltered and F0-shifted tokens. The factorial design for the analysis was thus a 2 (TALKERS) by 2 (DIRECTION of F0 shift) by 2 (processing METHOD) design nested within listeners. An additional analysis of variance was conducted with the raw percentage correct scores to provide information about absolute word identification rates for both the original tokens and the F0-shifted tokens.

B. Results

Absolute word identification accuracy was moderately high overall (an average of 72.5 percent of the words correct). However, listeners identified words correctly more often for the adult talker ($F[1,19] = 138.47$, $p < .01$) with 77.5 percent correct responses for the adult versus 67.4 percent correct for the child talker.

For the comparison of signal processing methods, we examined the difference between the percent words correct in the F0-shifted tokens and the original token. Overall, the signal processing was responsible for an increase of 3.7 in the percentage of errors. There was no significant main effect of pro-

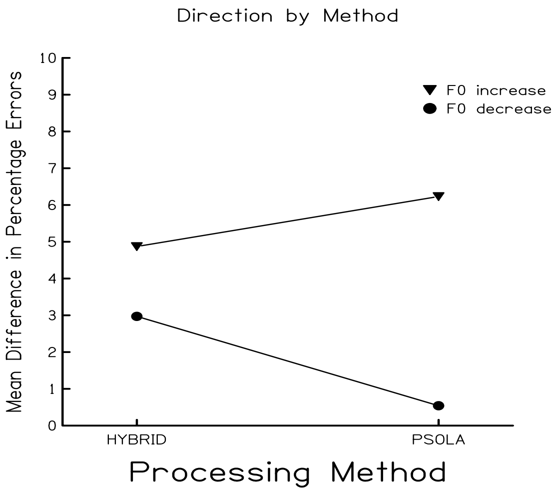

Fig. 2. Difference in percentage of identification errors for direction of F0 shift as a function of processing method.

cessing METHOD ($F[1,19] = 0.62$, $p = 0.44$). The main effect of F0 shift DIRECTION was significant ($F[1,19] = 24.86$, $p < .01$) with increased F0 resulting in an increase of 5.6 in percentage of errors, while decreasing F0 raised the percentage of errors by only 1.8. DIRECTION interacted significantly with METHOD ($F[1,19] = 5.37$, $p = .032$). This interaction is illustrated in Figure 2 which shows that the PSOLA method performed better than HYBRID when F0 was shifted down and worse than HYBRID when F0 was increased.

The main effect of TALKER was marginally significant ($F[1,19] = 4.41$, $p = .049$) with the child's speech being less distorted by F0 shift than the adult's (a change in percentage error of 2.3 for the child and 5.0 for the adult). The interaction METHOD and TALKER was also significant ($F[1,19] = 5.06$, $p = .037$). Means for this interaction are plotted in Figure 3 showing that the PSOLA method is again associated with both the best and worst performance.

No other interactions involving METHOD were significant, although the interaction of TALKER by DIRECTION of F0 shift was significant ($F[1,19] = 12.82$, $p < .01$) due to a disproportionately larger number of errors for the adult when her F0 was increased.

C. Discussion

Altering natural speech in any non-trivial way almost invariably degrades it. The present study compared two methods for altering F0 in natural speech which are intended to generate minimal signal degradation. Overall, both methods were found to succeed about equally well when intelligibility was used as the measure of signal degradation. De-

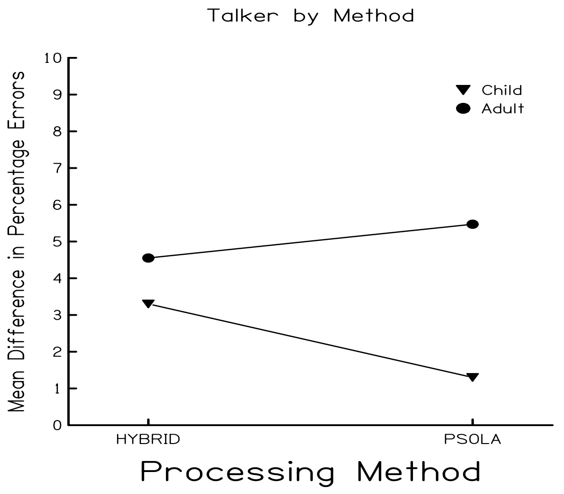

Fig. 3. Difference in percentage of identification errors for each talker as a function of processing method.

spite the gross similarity in effects on speech intelligibility of the two methods, several aspects of the results can be considered.

First, this study was motivated by the observation that the HYBRID method produced speech that was subjectively preferable to the PSOLA method. The basis of this subjective impression may be related to the observation in the present study that PSOLA tended to produce both the best and the poorest speech in the experiment. In subjectively assessing speech quality, listeners may assign greater weight to the perceptible failures of a signal processing method than to the (effectively imperceptible) successes of the method. This is at least consistent with our results which suggest that degradation from the HYBRID method may be less severe but more common than degradation due to PSOLA.

A second factor in comparing the two methods is stability across talkers. For the two talkers in the present study, we found that the adult's speech was more degraded by F0 shifts than was the child's speech. Note that overall, the adult's stimuli were *more* intelligible that those of the child talker, but they were also more adversely effected by shifting F0. This was true independent of processing method (although PSOLA produced slightly more degradation than HYBRID LPC for the adult talker) and independent of direction of F0 shift. While LPC-based methods are often cited as being heavily effected by individual talker differences, we see no evidence of this being a problem in the present case since the HYBRID method was slightly less effected by talker differences than was the PSOLA method.

Finally, we consider the finding in the present study that PSOLA performed better in lowering F0, while the HYBRID method performed better in raising F0 (cf. Figure 2). Recall that, with the HYBRID method, F0 is lowered by appending zeroes onto the residual to extend each pitch period. Thus, only information in the filter coefficients is retained during the extended portions of pitch periods, and that information may be unavailable if the filter ringing decays rapidly (i.e., if the poles are relatively broad in bandwidth). By contrast, with the PSOLA method, pitch period length must be more than doubled before the extended portion of each pitch period contains only zeroes. In effect, PSOLA retains more information about the original speech signal in the extended portions of each pitch period. If this is the case, the HYBRID method could be improved by (a) increasing the filter order to retain more information about the speech signal, (b) increasing the effective length of a pitch frame by windowing in the way PSOLA does, or (c) a combination of the two.

While the Hybrid LPC method we've described is not overall better at preserving speech intelligibility than the PSOLA method, it may produce slightly better sounding speech, and seems to have advantages for some talkers and/or when F0 is increased. Future studies will examine possible reasons for the poorer performance of the HYBRID method when F0 is decreased.

References

[1] C. Hamon, E. Moulines, and F. Charpentier. A diphone synthesis system based on time–domain prosodic modifications of speech. In *Proceedings of the 1989 IEEE International Conference on Acoustics, speech, and Signal Processing*, Glasgow, Scotland, 1989.

[2] M. Macchi, M. Altom, D. Kahn, S. Singhal, and M. Spiegel. Intelligibility as a function of speech coding method for template–based speech synthesis. In *Proceedings Eurospeech '93*, 1993.

[3] M. Spiegel, M. Altom, M. Maachi, and K. Wallace. Comprehensive assessment of telephone intelligibility of synthesized and natural speech. *Speech Communication*, 9(4), Aug. 1990.