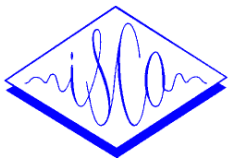

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

Intelligibility as a function of speech coding method for template-based speech synthesis

Marian Macchi, Mary Jo Altom, Dan Kahn, Sharad Singhal, Murray Spiegel

Bellcore (Bell Communications Research)
445 South Street, Room 2E-236, Morristown, New Jersey, USA 07960

ABSTRACT

We have been experimenting with various methods for coding the templates used in a concatenative speech synthesis system: standard pulse/noise-excited LPC; a newer waveform technique, PSOLA (pitch-synchronous overlap-and-add); and two types of residual-excited LPC (RELP): simple RELP, in which the residual was modified by truncation or padding with zeros, and PSOLA RELP, in which PSOLA was used to modify the residual. We used these techniques to code spoken words that were similar to the templates in an inventory, and resynthesized the words. We also modified the pitch of the words, as is required by text-to-speech synthesis systems, and resynthesized the pitch-modified words. We conducted listening tests to measure the consonant intelligibility in the words with and without the pitch change. Thus we were able to see how intelligibility was affected by the coding method itself and by changes to the pitch. The results showed that RELP provided higher intelligibility than PSOLA for voiced consonants, and considerably higher intelligibility than standard pulse/noise-excited LPC, even when pitch changes were imposed on the words. In addition, simple RELP performed about as well as PSOLA-RELP.

Keywords: *speech synthesis, speech coding*

1. INTRODUCTION

We have been experimenting with various methods for coding the templates used in a concatenative speech synthesis system. For concatenative synthesis, the templates are extracted from natural speech, coded with some type of speech coding method, and stored in an inventory. To produce a new utterance, the synthesis system concatenates the coded templates and modifies them in terms of pitch (F0), duration and perhaps spectrum, according to the linguistic environment of the new utterance. The goal of the experiments we describe here is to choose a speech coding technique for a new version of ORATOR®, Bellcore's concatenative text-to-speech synthesis system.

One of the basic requirements for a template-based speech synthesis system is that the speech be highly intelligible. Therefore, our goal is to choose a speech coding technique that has high intelligibility. The intelligibility of the speech produced by a synthesis system is in part a function of the speech coding technique itself, as can be seen in coding systems for transmission of human speech. However, because in a concatenative text-to-speech system the templates are modified in terms of pitch, duration, and spectrum, it is important to choose a coding method that preserves intelligibility when these changes are imposed on the templates.

Changes to pitch seem to have a significant effect on speech quality. Therefore, we have conducted listening experiments to evaluate, as a function of coding method and changes to the pitch, the intelligibility of consonants in words similar in structure to those from which the templates of the ORATOR inventory were extracted.

We have experimented with and evaluated four basic techniques: (1) standard, fixed frame LPC,[1] (2) pitch-synchronous LPC, (3) PSOLA (pitch-synchronous overlap and add),[2] [3] [4] a newer time-domain waveform technique that has been said to produce higher quality speech than LPC, and (4) two types of residual-excited LPC (RELP): simple RELP, in which the residual is truncated or padded with zeros[5] and PSOLA RELP, in which PSOLA is used to modify the residual.[3]

ORATOR is a trademark of Bellcore.

Coding Techniques

LPC-asynchronous: LPC analysis treats the speech signal according to source-filter theory. Asynchronous LPC analysis is performed at a fixed frame rate (in our case, 10 ms intervals) using a fixed window size (in our case, 25.6 ms with a Hamming window) and ignoring the voiced/unvoiced boundaries. We performed LPC analysis to compute 12 poles, corresponding to a maximum of six resonances, which is appropriate for 5kHz speech bandwidth (the bandwidth of our speech material). For synthesis, the excitation is a pulse train for voiced (periodic) portions of speech, or noise for unvoiced (aperiodic) portions of speech. LPC analysis is sensitive to the harmonic structure of the line spectrum - that is, LPC analysis often picks poles at peaks in the line spectrum of speech rather than the actual formant frequencies. However, it is the actual formant frequency that determines perception. Consequently, on resynthesis, LPC synthesis can reproduce a harmonic peak rather than the formant, so that it is possible that the resulting speech will be perceived as a different phoneme or as a distorted version of the intended phoneme.

In LPC synthesis, speech is synthesized using pulse excitation for the voiced frames and white noise excitation for the unvoiced frames. The time interval between pulses is determined by the desired pitch period. The LPC coefficients associated with the frame nearest to the start of the synthesis pitch period were used for synthesis.

LPC-synchronous: LPC analysis was performed on the speech pitch-synchronously (at intervals of 1 pitch period, with a Hamming window over 3 pitch periods).[6] Onsets of pitch periods were marked as peaks in the LPC residual, as described below ("*Speech analysis*"). The pitch (and consequently the frame rate) of the speech for this experiment ranged from about 75 Hz to 140 Hz. Unvoiced segments of speech were analyzed at a fixed frame rate (10 ms intervals) over a window of 25.6 ms. Synthesis was the same as for the LPC-asynchronous technique.

Residual-excited LPC (RELP): In the RELP method, LPC analysis is performed pitch-synchronously, as described above. The LPC residual is stored for use as the excitation signal, along with marks indicating pitch period onsets. For voiced frames, the residual must be modified when the pitch is changed, but the residual for frames of unvoiced speech is not modified.

Simple RELP: For simple RELP, when the pitch is raised, the LPC residual is truncated at the end of the new, shorter pitch period; when the pitch is lowered, the residual is extended with zeros. To minimize discontinuities near the peak of the LPC residual, the pitch period was defined to begin 5 samples (.5 ms) earlier than the marked pitch period onsets.

PSOLA RELP: For pitch changes in PSOLA RELP, the LPC residual is modified by the PSOLA technique described below. The same window-selection strategy was used as for PSOLA, and the resulting signal was used as the excitation signal for LPC synthesis.

PSOLA: The PSOLA technique is not, strictly speaking, a speech coding technique. It does not decompose the speech signal spectrally - it operates directly in the time domain on the speech waveform and does not parameterize the signal. It is extremely simple computationally. PSOLA requires that the voiced (periodic) portions of the speech waveform be segmented into pitch periods and weighted. In our case, a Hanning window was used; the window was centered on the beginning of the pitch period as defined by our algorithm that marked period onsets at peaks in the LPC residual. To raise the pitch from that of the original speech (synthesize speech with a shorter period than the original), the weighted pitch periods are overlapped in time; to lower the pitch (synthesize speech with a longer period than the original), the weighted pitch periods are spread further apart in time. To derive the new speech, the weighted periods are added and divided by the sum of the weights. To raise the pitch we used the new, shorter pitch period as the weighting window. To lower the pitch we used the original period as the weighting window. Unvoiced portions of speech are unaffected by pitch changes.

2. EXPERIMENTAL DESIGN

A listening experiment measured the intelligibility of consonants in words similar to those from which the templates for our speech synthesis inventory were extracted. We chose to measure intelligibility of consonants, because we assumed that consonant intelligibility would be more sensitive to coding method and pitch changes than would vowel intelligibility. The words were recorded, then analyzed and resynthesized, with and without a change to the pitch, with each of the coding techniques. The words were 156 monosyllabic nonsense words from the Bellcore Monosyllabic Corpus, which includes most consonant clusters in English, spoken in isolation with a declarative intonation pattern by a male speaker.[7]

Speech Analysis

The speech was low-pass filtered at 4.5kHz and sampled at 10ksps. The voiced/unvoiced decisions were done by hand, with the criterion that if a segment indicated the presence of any periodic voicing, the segment should be considered voiced. For the pitch-synchronous methods (pitch-synchronous LPC, PSOLA, and PSOLA-residual-excited LPC), the onsets of the pitch periods in the speech were marked. We assumed that the onset of a pitch period - the instant of main vocal tract excitation - would be represented by a peak in the LPC residual. However, peak-picking in a continuous function like the LPC residual can be difficult. We expected that the pitch period onset would occur near a pulse in the multipulse signal, so we developed an algorithm to use the pulses in a multipulse excitation signal as a guidepost to pitch period onsets in the LPC residual. For the speech used in this experiment, the average pitch period was around 10 ms. Consequently, we performed asynchronous LPC analysis to compute two multipulses every 10 ms. The pulses in the multipulse excitation signal plus heuristics to exclude pitch halving and doubling were used to determine pitch period onsets in the residual. Errors in marking pitch period onsets were hand-corrected.

Pitch Modifications

The words were synthesized in two ways: (1) without any change in pitch, so that essentially nothing more was done to the speech than analyzing and resynthesizing it; (2) with a rising pitch pattern copied from a natural utterance of the word with a yes-no question intonation pattern. Our reason for doing synthesis both with and without pitch modification was to determine the magnitude of degradation of intelligibility due to the coding technique itself versus that due to the effects of changing the pitch.

We chose to use a rising pitch pattern for the pitch-altered stimuli because it introduced a relatively large pitch change to the templates. The rising pitch pattern rose from 66 Hz to 142 Hz. In the original utterances, the average F0, across all words, at the beginning of voicing was 102 Hz; the average at the peak of the F0 was 116 Hz, and the F0 pattern fell to an average F0 of 75 Hz at the ends of utterances.

For each corpus word, the change in pitch was produced by linearly stretching or compressing the rising pitch pattern to fit the duration of the voiced portion of the word. The rising pitch pattern was *not* varied to incorporate micromelody, the differences in F0 that are associated with the phonetic identity of the segments in the syllable.

Experimental Procedure

We conducted a formal listening test with the coded words. Listeners were college students who were paid for their participation. Each listener heard only one coding technique, and each listener heard either the non-pitch-changed version or the pitch-changed version of the words. There were nine listeners for each of the conditions. The words were played to listeners over telephone bandwidth (250-3300 Hz) through headphones in a sound-treated booth. The listeners' task was to write down the syllable they heard. The listeners' transcriptions were scored by hand[7] and represented as intelligibility scores (the percentage of consonants that were correctly transcribed). Intelligibility scores were analyzed via analysis of variance, and comparisons between the conditions were performed with the Duncan new multiple range test (adopting $p < .05$)[8] using subjects (9/condition) as the error term.

3. RESULTS

Table 1 gives means and standard errors for the overall intelligibility scores (N/cell = 1404 = 9 listeners x 156 words).

Table 1
Overall Intelligibility Scores

Coding	No Pitch Change		Pitch Change	
	Mean	SE	Mean	SE
Natural	94.4	0.6	—	
Simple-RELP	91.9	0.7	92.0	0.7
PSOLA-RELP	91.9	0.7	92.7	0.7
PSOLA	—	—	90.2	0.7
LPC-Synch	86.1	0.9	79.3	1.1
LPC-Asynch	83.6	1.0	81.7	1.1

As shown in Table 1, intelligibility ranged from 79.3% for LPC-synchronous, with pitch change, to 94.4% for natural speech. The residual-excited LPC techniques - both the simple RELP and the PSOLA RELP - had the highest intelligibility of the techniques, and not significantly different from natural speech (the intelligibility in the pitch-change condition for PSOLA RELP was only 1.7% lower than natural speech, and for simple RELP, 2.4% lower than natural speech). The intelligibility of PSOLA was significantly lower than natural speech (4.2% lower), and slightly, but not significantly, lower than the RELP techniques (1.8% lower than simple RELP) and (2.5% lower than PSOLA RELP). The other LPC techniques showed considerably lower intelligibility than natural speech

and the other coding techniques, with a large component due to the coding method itself, as well as to the pitch change. Pitch-synchronous LPC did exhibit a small advantage over asynchronous LPC when pitch was not changed but any advantage was not preserved in the pitch-changed condition.

Table 2 shows the scores for voiceless consonants (N/cell = 801 = 89 words with voiced consonants x 9 listeners). Table 3 shows the scores for voiced consonants (N/cell = 603 = 67 words with voiced consonants x 9 listeners).

Table 2
Intelligibility Scores - Voiceless Consonants

Coding	No Pitch Change		Pitch Change	
	Mean	SE	Mean	SE
Natural	94.9	0.8	—	—
Simple-RELP	93.0	0.9	92.1	1.0
PSOLA-RELP	93.0	0.9	92.3	0.9
PSOLA	—	—	92.6	0.9
LPC-Synch	86.1	1.2	78.7	1.5
LPC-Asynch	84.1	1.3	82.9	1.4

Table 3
Intelligibility Scores - Voiced Consonants

Coding	No Pitch Change		Pitch Change	
	Mean	SE	Mean	SE
Natural	93.9	1.0	—	—
Simple-RELP	90.4	1.2	91.9	1.1
PSOLA-RELP	90.4	1.2	93.4	1.0
PSOLA	—	—	87.1	1.4
LPC-Synch	86.1	1.4	80.1	1.6
LPC-Asynch	82.8	1.5	80.0	1.7

For voiced consonants, compared to voiceless consonants, place of articulation is cued more by voiced portions of speech. If consonant intelligibility is decreased by degradation of the acoustic cues local to the consonant, we might expect voiced consonants to suffer more degradation than voiceless consonants in the pitch-changed words, since only the voiced portions of the signal were modified as a function of the pitch change.

In fact, for voiceless consonants, the RELP techniques and PSOLA all had intelligibility scores near, and not significantly different from, natural speech, even in the pitch-changed condition. For voiced consonants, the RELP techniques also had intelligibility scores near, and not significantly different from, natural speech. Further, there was no significant difference in intelligibility between the simple RELP and the PSOLA-RELP for either voiced or voiceless consonants.

In contrast, for PSOLA the voiced consonants had significantly lower scores than the RELP techniques for the pitch-changed condition (4.8% lower than simple RELP and 6.3% lower than PSOLA-RELP) and natural speech (6.8% lower). In addition, the intelligibility of PSOLA-coded voiced consonants was significantly (5.5%) lower than that of PSOLA-coded voiceless consonants. From this result we infer that the pitch modifications performed with PSOLA degrade voiced consonant intelligibility more than do the same pitch modifications with the RELP techniques.

There was no difference in intelligibility between voiced and voiceless consonants for pitch-synchronous LPC, but voiceless consonants had higher intelligibility than voiced consonants for pitch-asynchronous LPC. We have no explanation for why the voiceless consonants had higher intelligibility in one type of LPC but not the other.

4. CONCLUSION

This study found that RELP provides higher consonant intelligibility for synthesized speech than PSOLA for voiced consonants, and considerably higher intelligibility than more standard pulse/noise excited LPC, in general, even when pitch changes are imposed on templates. The advantage was particularly salient for voiced consonants, whose acoustic-phonetic cues are more affected by pitch changes than are those for voiceless consonants. In addition, simple RELP performed about as well as PSOLA-RELP.

REFERENCES

[1] Atal, B. S. and Hanauer, S. L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoustical Society of America, 50(2): 637-655.

[2] Hamon, C., Moulines, E. and Charpentier, F. (1989). "A diphone synthesis system based on time-domain prosodic modifications of speech," ICASSP, p. 238-241.

[3] Moulines, E. and Charpentier, F. (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, p. 453-467.

[4] Valbret, H., Moulines, E., and Tubach, J. P. (1992), "Voice transformation using PSOLA technique," ICASSP, p. I145-I148.

[5] Caspers, B. E. and Atal, B. S. (1983). "Changing pitch and duration in LPC synthesized speech using multipulse excitation," J. Acoustical Soc. Amer, suppl1., vol 73, p. S5.

[6] Singhal, S. and Atal, B. S. (1984), "Improving performance of multi-pulse LPC coders at low bit rates," ICASSP, pp. 1.3.1-1.3.4.

[7] Spiegel, M., Altom, M. J., Macchi, M. J., and Wallace, K. (1991), "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech," Speech Communication, 9, 279-292, 1990.

[8] Kirk, R. E. (1968), *Experimental design: procedures for the behavioral sciences.* Wadsworth Publishing Company, Belmont, California.