SPEECH SYNTHESIS USING THE CELP ALGORITHM

Geraldo Lino de Campos

Evandro Bacci Gouvêa

Escola Politécnica da Universidade de São Paulo

Dept. de Engenharia de Computação

P. O. Box 8174 - São Paulo 01065-970 - BRAZIL

Phone + 55 11 818 5288 - FAX + 55 11 818 5718

e-mail: geraldo@regulus.pcs.usp.br

ABSTRACT

This paper presents a phoneme/diphone based speech synthesis system for the (Brazilian) Portuguese language. The basic idea bearing this system is the construction of a library of phonetic units, and processing of those basic units to build an utterance. The system is complemented by a text to phoneme translator described in [Cam95].

The phoneme's representation in the library is based on a linear prediction model; the filter which models the vocal tract is represented by Line Spectrum pairs, and the excitation by Code Excited Linear Prediction (CELP) parameters.

Thus paper is organized as follows. After a brief introduction, CELP coding is briefly presented in part 2. Part 3 presents the relevant points to be applied in speech synthesis. Parts 4 and 5 constitutes the main contribution of this paper, detailing the process of building the phoneme library and the interpolation techniques used. Part 6 presents some concluding remarks.

1. INTRODUCTION

CELP (Code Excited Linear Prediction) is a technique developed for speech compression. However, it can be extended to voice synthesis, with results far superior than other usual techniques. This paper presents problems and results of CELP usage in voice synthesis for the (Brazilian) Portuguese language. The speech produced is of very good quality, almost natural, in a level of quality unattainable by other synthesizers based on LPA (Linear Predictive Analysis).

Systems based on LPA usually deliver a very good representation of the vocalic tract resonances; the problem is with the representation of excitation, usually constrained to a simple composition of pulses and white noise. CELP addresses this point, using an elaborated model of excitation.

2. CELP CODING

CELP was recently developed [Cam89, Lin86, Cox89], and has been adopted as an US Federal Standard [Fed91]. It uses the usual LPA for determining the resonance of the vocal tract; its main innovation is the way to handle the excitation.

Very briefly, CELP coding is based on analysis-by-synthesis search procedures. First, a standard LPA is performed, resulting in a 10th order filter. The filter parameters are transmitted in the form of LSP (Line Spectrum Pairs), an equivalent representation with better quantization properties, and are called the LSP parameters. A set of different excitations are applied to the filter. Each output is compared with the input signal, using a perceptually weighted distance measure. The excitation with the least distance is selected and transmitted. Obviously, it is necessary to reduce the possible excitations to a number as small as possible. This is accomplished by a fixed stochastic code book for representing the unvoiced components of speech, and an adaptive code book for the voiced components. The information transmitted are the indexes in these code books, with their respective gains, called the CELP parameters.

Although the computational requirements are high, requiring a digital signal processor with 25 MIPs for real-time operation, such processors are readily available and of low cost. In [Det92], a figure of US$200.00 is quoted as a typical value for a one channel CELP system, if produced in small quantities.

CELP coding systems produce a speech that is almost indistinguishable from natural speech, when recorded in an office environment. As it uses a larger bandwidth, 3.7 instead of 3.2 KHz, the perceived quality is better than the obtainable with the usual phone system. More important, it is very robust to environment noise, a feature that older systems lacked dramatically.

3. APPLICATION TO SPEECH SYNTHESIS

The speech synthesizer described here is part of a text-to-speech system; text is first converted to a sequence of phonemes [Cam95], pitch information is added and the voice synthesizer is invoked.

The natural way to implement a voice synthesis system is by the construction of a library of basic sound units. It is possible to use only phonemes, and compose all utterances by further processing. However, the composition of some sequences may be very difficult to achieve, due in part to coarticulation effects and in part to the fact that the usual interval analysis , of the order of 20 ms, is too large to capture effectively the behavior of phonemes with short duration - the plosives, for instance. Results can be much better using diphones in these situations, and our library is composed of both phonemes and diphones, as described in [Cam85].

After the CELP analysis of an utterance containing the desired phoneme or diphone, we obtain LSP parameters representing the configuration of the vocalic tract, and CELP parameters representing the excitation. Both are kept in the library, and must be processed in order to produce an utterance. LSP parameters can be processed as usual, and will not be considered further. The important issues are generation and control of gain and pitch, i.e., the CELP parameters.

The adaptive code book contains a history of past excitations. During the analysis of the first samples, this code book contains incorrect information, that cannot be used for synthesis. This can be circumvented by zeroing the gain for the adaptive code book during a suitable amount of time at the beginning of an utterance. By zeroing also the gain of the stochastic code book, it is possible to obtain a null excitation signal, forcing the zeroing of adaptive code book contents for the samples in question. Extending these ideas for a synthesizer, it is necessary to repeat these steps for every situation when a signal with non-zero amplitude must be generated. This causes the only aspect that compromises the naturalness of the generated speech: utterances starting with strong voiced sounds are slightly different than natural sound.

Pitch can be controlled by manipulating indexes in the adaptive code book. In CELP, the adaptive code book index indicates the past excitation that better represents the current excitation. If the pitch were constant, this index will be a direct measure of the pitch period. Changing the index value will change the pitch accordingly, resulting in a very natural pitch sounding.

Gain processing is somewhat more involved. Considering an interval with no stochastic excitation, the adaptive code book gain multiplies a vector that represents one of the past excitations of the system. As such, a value of 1 means a constant level; changing levels is essentially a change of gain during one analysis (or synthesis) interval. The processing of gain values is critical in the sense that they are very susceptible to noise or processing error.

In the general case, when both adaptive and stochastic gains are present, sometimes the reduction of one gain is compensated by the increase of the other, and the manipulation algorithms must be well elaborated to assure good and uniform results.

4. PHONEME LIBRARY

The construction of the phoneme library follows the ideas described in [Cam85]. Utterances containing the diphones of interest were recorded, and the segments containing the diphones were selected.

Since Portuguese is essentially a syllabic language, the majority of the diphones are of the form <consonant> <vowel>. One word containing the diphone was selected, according to the following criteria, adopted to minimize coarticulation effects:

- the utterance must be an existing word in the language;

- the diphone must be the stressed syllable of the word;

- the diphone must not be the first or last syllable of the word;

- the consonant must be preceded by a vowel;

- the syllable following the phoneme should start with a plosive, if at all possible.

These criteria allows both an easy isolation of the diphone and no interference of coarticulation effects.

It was found that is very important to use real words, since isolated syllables are uttered in a quite different and unnatural way. With meaningless word the situation is a little better, but they should be avoided by the same reason.

Stressed syllables were selected because they are usually uttered with far more clarity then the other syllables in the word. We do not know of formal studies about this fact, but it seems to be fairly general, and not a specific property of the spoken Portuguese language.

The chosen word for recording each diphone was also required to have a plosive sound after the selected diphone to simplify the identification delimitation of its last frame.

To improve the characterization of individual consonants, it was further required that the consonant was bounded by vowels. Vocalic sound are easily recognized in the spectrogram, defining clearly and unambiguously the limits of the consonant.

Words containing the semivowels (/l/ and /r/ in Portuguese) were selected following essentially the same criteria, and included in the library.

The steps required for each phoneme/diphone were

- recording of the word;

- analysis by the standard CELP algorithm;

- isolation of the phoneme/diphone by visual inspection of a spectrogram generated from the LSP parameters;

- the duration was standardized to 195 ms (26 7.5 ms frames);

- synthesis of the selected phoneme/diphone, for checking purposes;

- inclusion in the library.

A further step is required after the library is assembled. Since the samples were taken from different contexts, they might have quite different intensities, and two normalizations are required. The first is intra-vowel intensity normalization.

At least in Portuguese, the intensity of a diphone depends essentially on the vowel intensity. So. diphones of the form <consonant><vowel> should have approximately the same intensity for the same vowel. All the diphones containing the same vowel were normalized for a the same intensity, by multiplying the stochastic gain by a suitable value. This value can be, for instance, the relation between the average original intensity and the desired normalized value.

It is further required to normalize inter-vowel intensities. That normalization is essentially subjective. For each two vowels, diphones with the same consonant are chosen; one is maintained constant, and the intensity of the second is varied until both are perceived with the same subjective intensity. All the diphones containing the second vowel have their intensity multiplied by that ratio. The procedure is repeated for all the vowels present in the library.

5. INTERPOLATION TECHNIQUES

Once the diphone library is available, the next problem is the composition of a given sequence of phonemes. The following techniques are heuristics that offer good results for the Portuguese language, and can be considered a starting point for developing a similar system for other languages.

In this section the synthesis will be considered left to right; i. e. always between an already processed sequence and an arriving diphone.

The Linear Prediction model generates voice by the excitation of a filter by an appropriate signal. Accordingly, the following discussion will show the interaction, in the phoneme boundaries, of both the parameters that represent the filter (LSP parameters) and the excitation parameters (CELP parameters).

There are three possibilities, depending on the nature of the incoming phoneme/diphone:

- the diphone starts with a plosive consonant;
- the diphone starts with another kind of consonant;
- the phoneme is a vowel.

The simplest case is a diphone starting with a plosive. Since there is an interval of silence separating the diphone from the preceding sound, there is no significant coarticulation effects, and no interpolation is necessary. The silence period preceding the start of the plosive sound is fixed in 30 ms. Although this interval varies in real speech, the adoption of this value allows avoiding coarticulation effects.

Initially, a silence interval was introduced. Although the speech was of good quality, the subjective feeling was strange for some listeners. To improve the synthesis quality, another strategy was tested, and removed that unpleasantness. The LSP coefficients were inter-

polated between the last interval of the previous diphone and the diphone being introduced. The pitch of the last interval of the previous diphone was maintained, and the intensity strongly reduced, by using an adaptive gain well below 1.

Encounters between a diphone starting by a non-plosive consonant are subject to strong coarticulation effects. The coarticulation according to [Sha87] tends to be a phenomenon that precedes the second diphone, as if the vocal tract was preparing for the new geometry required for the incoming phoneme, using the last and less important part of the preceding diphone.

To simulate this effect, the LSP parameters were interpolated for a variable interval involving mainly the last part of the previous diphone. The best results, based on subjective sound quality, were obtained by using a 60 ms interval, starting 45 ms before the end of the previous phoneme. The LSP parameters were linearly interpolated; the CELP parameters didn't require any processing. The solution was simpler than expected.

Processing of vocalic encounters is critical in Portuguese. Although the crudest concatenation generates understandable utterances, even small deviations from the natural sounds are easily identifiable and quite annoying.

Sharp transitions are rare; usually, diphthongs are generated. As diphthongs correspond to slow transitions of the vocalic tract, the resonance aspect can be easily simulated by interpolation of the LSP parameters. It was found experimentally that a duration around 160
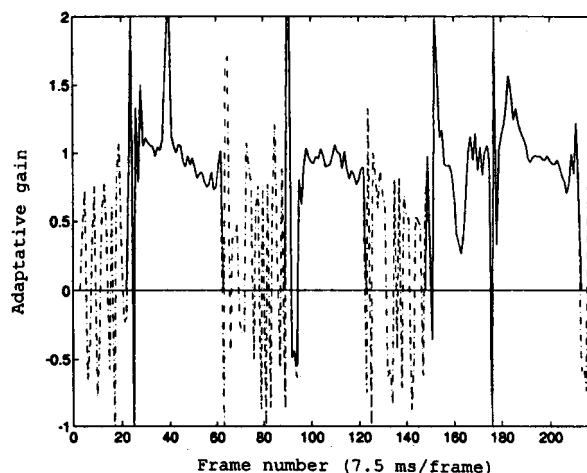

Figure 1 - Adaptative gain for an utterance containig three diphtongs, showing the regular decay pattern.

ms, symmetrically distributed in the phonemes offered the best results for vowel-vowel encounters, and 80 ms for encounters involving semi-vowels.

Processing of the CELP parameters is somewhat more complex. The profile of the CELP parameters of a diphthong has the same general shape of the parameters for a single vowel: the adaptive index tend to remain constant, the adaptive gain tends to 1, and the stochastic gain decreases. This parameter profile must be generated

over the total length of the synthesized diphthong, starting at 1.0 and going to 0.75. Figure 1 is a plot of the adaptative gain for a natural utterance containing three diphtongs, showing the regularity of the patern.

This is not enough, however. The subjective intensity of each vowel is different, and the direct application of the above rule must be changed to take in account that difference. The value of the adaptive gain must be changed to the inverse ratio of subjective intensity of each vowel.

6. CONCLUSIONS

CELP coding techniques offer a strong basis for speech synthesizer systems. The quality of synthetical voice is far superior to other techniques using LPA. Algorithms are relatively simple and straightforward.

CELP is only one of a family of analysis-by-synthesis systems; since they differ mainly in coding details, it is expected that any of them will be a suitable basis for developing voice synthesizers.

The experiments were conducted for the Portuguese language, that has very few consonant encounters to be synhtesised. The results suggest that the technique can be easily extended for other languages.

7. REFERENCES

Cam85 Campos, Geraldo L. Speech synthesis for unrestricted vocabulary: the phoneme-diphone approach. Proceedings of the Speech Tech' 85, New York, New York, 307-309, 1985.

Cam89 Campbell, J.,Welch, V. and Tremain, T. "An Expandable Error-Protected 4800 bps CELP Coder", Proc. of ICASSP, 735-738, 1989

Cam95 Campos, Geraldo L. and Chbane, Dimas T. "A Finate Automaton Translator of Text to Phonemes for the Portuguese Language". Annals of the XIIIth International Congress of Phonetic Sciences, 28.2-8, 1995.

Cox89 Cox, R., Kleinj, W. and Kroon, P. "Robust CELP Coders for Noisy Background and Noisy Channels", Proc. of ICASSP, 739-742, 1989.

Det92 Details to Assist in Implementation of Federal Standard 1016 CELP. Office of the Manager, National Communications System, Jan. 1992.

Fed91 Federal Standard 1016 - Telecommunications: Analog to Digital Conversion of Radio Voice by 4800 bit/second Code Excited Linear Prediction (CELP), National Communications System, Office of Technology & Standards, Washington, DC 20305-2010, Feb. 1991.

Lin86 Lin, D. "New Approaches to Stochastic Coding of Speech Sources av Very Low Bit Rates", Signal Processing III: Theories and Applications, Proc. of EUSIPCO-86, 445-448, 1986.

Sha87 O'Shaughnessy, D. "Speech communication: human and machine", Addison-Wesley, New York, 1987.