Towards the Speech Synthesis of Raramuri: A Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences

Alfonso Medina Urrea,[1] José Abel Herrera Camacho[2]
and Maribel Alvarado García[3]

[1] GIL-II, Universidad Nacional Autónoma de México
04510 Coyoacán, DF, MEXICO
amedinau@ii.unam.mx
[2] FI, Universidad Nacional Autónoma de México
04510 Coyoacán, DF, MEXICO
abelh@verona.fi-p.unam.mx
[3] Escuela Nacional de Antropología e Historia
14030 Tlalpan, DF, MEXICO
marvarado1978@yahoo.com.mx

Abstract. This work deals with the design of a synthesis system to provide an audio database for Raramuri or Tarahumara, a Yuto-Nahua language spoken in Northern Mexico. In order to achieve the most natural speech possible, the synthesis system is proposed which uses a unit selection approach based on function words, suffix sequences (derivational and inflectional morphemes) and diphones of the language. In essence, the unknown suffix units were extracted from a corpus and recorded, along diphones and function words, in order to build the audio database that provides data for Text-to-Speech synthesis.

1 Introduction

The ultimate objective of Text-to-Speech (TTS) synthesis systems is to create applications which listeners, and users in general, cannot easily determine whether the speech he or she is hearing comes from a human or a synthesizer.

Synthesized speech can be produced by concatenating recorded units (waveforms) selected from a large, single-speaker speech database. The primary motivation for using a database with a large number of units that covers wider prosodic and spectral characteristics, gives us the great benefit to produce a synthesized speech that sounds more natural than those produced by systems that use a small set of controlled units (*e.g.* diphones) [1]. There is a paradigm for achieving high-quality synthesis that uses a large corpus of recorded speech units; it is called *unit-selection synthesis*. Unit selection is a method in which we can concatenate waveforms from different linguistic structures such as sentences, words, syllables, triphones, diphones and phones. Due to the increasing computer's storage capacity, we are able to create a corpus of prerecorded

Advances in Computational Linguistics.
Research in Computing Science 41, 2009, pp. 243-256

Received 15/11/08
Accepted 11/12/08
Final version 05/02/09

units. Furthermore, there are efficient searching techniques that allow a real-time searching into huge databases looking for sequences of units in order to build up the synthesized utterance.

The objective of unit selection systems is to search an audio database in order to find the optimal sequence that makes up a target utterance. The unit selection is based on minimal acoustic distortions (cost) between selected units and the target spectrum [2]. As Zhao establishes [3], "the cost function measures the distortion of the synthesized utterance; this is a summation of two sub-cost functions: a target cost, which describes the difference between the target segment and the candidate segment, and a concatenation cost, which reflects the smoothness of the concatenation between selected segments."

We are motivated to work with the Raramuri language group, which is constituted by a cluster of five of variants, because it is one of the relatively least endangered groups. It is worth noting that 364 variants, belonging to 68 language groups and 11 linguistic families, have been recognized rather recently as official[4] proper languages of Mexico; a true linguistic continent. Certainly, such linguistic wealth deserves to be studied in order to develop technologies that so far have been considered necessary only for the dominant languages of the world. And Raramuri seems a good place to start with. Regarding its relatively unendangered status, although government statistics are subject to question, in 1970 more than 25 thousand Raramuri speakers were counted, whereas today around 75 thousand speakers are estimated. Also, the phonological resemblance between Raramuri and Spanish and the restricted syllable structure of the former (CV) are additional motivations for our team to work with Raramuri; especially this last point has a positive impact on our TTS approach. The main challenge is that the language is not sufficiently known in order to be able to find somewhere in the bibliography enough data about the units to be used in the system we propose. This should illustrate the importance of conducting basic linguistic research in order to develop language technologies, since it is no secret that most world languages are not sufficiently documented.

2 Synthesizer

TTS is defined as "the production of speech by machines, by way of the automatic phonetization of the sentences to utter" [4]. The two characteristics used to describe the quality of a speech synthesis system are naturalness and intelligibility. The most common methods for speech synthesis are: articulatory, formant, and concatenative synthesis. Nowadays, the last two are the most used methods [5]. The formant synthesis have produced the most natural voice. However these systems provide excellent quality for some phrases and robotic voice for others [6, 7]. Our experience is that concatenative methods are more consistent than the formant or sub-phoneme ones.

[4] Instituto Nacional de Lenguas Indígenas, *Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*, Mexico, http://www.inali.gob.mx/catalogo2007/.

In concatenative systems, segments of prerecorded speech are chained together to form words, phrases and so on. These methods provide not much naturalness to the output speech, and it also results in audible glitches. However several methods based on overlapping adds have been applied to provide naturalness [8]. Also, it is usual to smooth the output waveform at the point of concatenation in order to gain naturalness. Three main sub-methods are used, based on the type of unit:

- unit-selection
- diphones
- domain-specific synthesis

Regarding concatenative synthesis, domain-specific synthesis requires a corpus of prerecorded words and phrases. It works well for systems with a very limited vocabulary (*e.g.* talking clocks). Also, a diphone is a segment that contains the stable parts of two adjacent phones; according to the specific phonotactics of the target language, the number of potential diphones for a given language is the square of its phones. Some must be discarded because they are incompatible with the phonotactics of the language. The exclusive use of diphones results in relatively small speech database, but the lack of clarity with the resulting speech is a disadvantage. Unit-selection requires a larger database; its corpus typically includes diphones, syllables, words, phrases and sentences. The synthesizer determines in real-time the best units from the database for the output utterance.

One of the major problems with all concatenative systems is how to deal with the boundaries between segments. It is clear that minimizing the number of occurrences of boundaries is likely to improve the quality of speech; reducing the number of boundaries involves, of course, using longer units. The point is: the longer the unit, the greater the number and detail of boundaries within them [9].

In theory, at the phone level there has to be an entry for every possible combination of phones and of phones and silence. But there are a number of combinations that do not exist in Raramuri and can be excluded. The original speech recording needs to be as monotonous as possible to reduce discontinuities between different segments and to reduce as much as possible any need for signal processing. This database must be stored in uncompressed PCM format to reduce compression-induced degradation of the signal. Once we have a stream of diphones the last step is to join them into a complete utterance.

The quality of concatenative synthesizers highly depends on the quality of the recorded speech units. Because of the speech sounds range from x to y, the speech corpus will be recorded in WAV format, under a sampling rate of 22,050 kHz with a 16-bit resolution. The recording sessions will be made in a professional studio. However, the recordings will be reduced later to a band of 4 kHz.

There are different types of units to be recorded. As mentioned above, we aim at function words, affix sequences and diphones. These segments were obtained from recorded utterances, preserving their suprasegmental features. In

our system, we use the three units forming a concatenative combinational system. These systems have better performance [9]; and the software designed to decide what unit will be used, and to query the increased database, does not pose a great challenge.

Another reason to use diphones as our basic units is that they model the joints well most of the time. Despite the price of storing a large number of permutations of units, this choice is much more convenient than using syllables because they would need a considerably larger number of permutations.

At first sight, it might be suggested that "the longer the unit length, the less troublesome will be any errors because larger 'amounts' of semantic content will be captured by each unit. If this is the case then errors in conjoining small units like phones will be the most critical for perception. This is precisely the reason why some researchers prefer diphones as the small units rather than phones —the very structure of the model is designed to minimize listener awareness of error, and errors are most likely to occur at coarticulated joins between segments." [9].

The idea of using larger units is simple, since coarticulation problems happen between individual units, for example between phones; The larger the unit, the greater the number of joints that will not require postprocessing. Therefore, whole function words and suffix sequences will be selected for recording, taking the most frequent sequences of function words found in the corpus.

Here, we give a brief description of the units involved in the system, ordered hierarchically from the lexical level to the phonetic one:

Function Words. Words are syntactic units; in this case, words are obtained from phrases. As it would be expected of any language in general, function words are the most frequent ones in the Raramuri corpus. The least frequent ones are normally content words which may be constructed by means of diphone concatenations.

Affix Sequences. Raramuri exhibits a small set of inflectional suffixes which tends to follow an interesting set of derivational suffixes.

Diphones. Diphones are defined as a stretch from the least varying (most stable or steady-state) part of a phone to a similar point in the next phone. The idea of introducing diphones was to capture the transition between phones within the acoustic model in order to reduce mismatches between phones.

Units which stand higher in this hierarchy already have internal boundaries modelled correctly by definition. That is, when they are used for concatenation, their suprasegmental features are implied.

The task of labeling consists of analyzing waveforms and spectrograms as well as making annotations to the waveforms of the recorded speech in order to extract information about the recorded units. In general, unit selection systems require phonetic labeling to identify limits between segments (phrases, words, diphones). It is also necessary to apply prosodic labels to give us information about tone and stress. Phrasal labels identify limits of each phrase recorded in the corpus. Word labels consist in time markers at the beginning and end of words. Tone labels are symbolic representations of the melody of the utterance. This

job is usually done by automatic speech labeling tools because of the database's size. For phonetic labeling, speech recognizers are used in forced alignment mode, where the recognizer finds the boundaries between segments. Automatic prosodic labeling tools work from a set of linguistically motivated acoustic features (e.g., normalized durations, maximum/average pitch ratios) plus some binary features looked up in the lexicon (e.g., word-final vs. word-initial stress) [10]. Unfortunately we don't have this kind of tools and the development of them would requiere more time to mark up text; because of this, we marked up the recorded units manually.

3 Identifying Units for Raramuri

Raramuri or Ralamuli, also known as Tarahumara, is a Yuto-Nahua or Uto-Aztecan language spoken in northern Mexico. It is more an agglutinative language than a fusional one. Word formation is mainly accomplished by means of suffixation. As could be expected, stems are followed by derivational suffixes, and these by inflectional ones. Also, its syllable structure is mainly CV, although syllable V is possible. Since unit-selection synthesis presupposes the units of the language that must be recorded in order to compile the database that will be used to build up the synthesized utterance, these facts are relevant in order to pick the appropriate units for the synthesizer. Hence, given the predominant

Table 1. Possible CV Syllables of Raramuri

CV	**a**	**e**	**i**	**o**	**u**	**a'**	**e'**	**i'**	**o'**	**u'**
m	ma	me	mi	mo	mu	ma'	me	mi'	mo'	mu'
n	na	ne	ni	no	un	na'	ne'	ni'	no'	un'
k	ka	ke	ki	ko	ku	ka'	ke'	ki'	ko'	ku'
p	pa	pe	pi	po	pu	pa'	pe'	pi'	po'	pu'
t	ta	te	ti	to	tu	ta'	te'	ti'	to'	tu'
c	ca	ce	ci	co	cu	ca'	ce'	ci'	co'	cu'
g	ga	ge	gi	go	gu	ga'	ge'	gi'	go'	gu'
b	ba	be	bi	bo	bu	ba	be'	bi'	bo'	bu'
r	ra	re	ri	ro	ru	ra'	re'	ri'	ro'	ru'
h	ha	he	hi	ho	hu	ha'	he'	hi'	ho'	hu'
w	wa	we	wi	wo	wu	wa'	we'	wi'	wo'	wu
y	ya	ye	yi	yo	yu	ya'	ye'	yi'	yo'	yu'
s	sa	se	si	so	su	sa'	se'	si'	so'	su'
l (R)	la	le	li	lo	lu	la'	le'	li'	lo'	lu'
	La	Le	li	Lo	Lu	La'	Le'	li'	Lo'	Lu'

syllable structure, it makes sense to pick diphones as the basic units. Table 1 shows the possible CV diphones according to the phonotactics of the language. Additionally, there are 50 VV diphones (two syllables) possible in the language.

Another obviously important kind of unit consists of the most frequent graphical words, which normally correspond to function words: pronouns, determiners, pospositions (instead of prepositions), conjunctions, prominent adverbs, frequent nouns and adjectives (numbers, colors and kinship words). Some of the 98 of these items appear in Table 2.

By including function words in the database, a synthesizer can be developed with relative fewer distortions than one using merely diphones. Thus, these latter would be used to build non function words, *i.e.* content words. Content words in Raramuri, as mentioned above, exhibit suffix sequences of derivational and inflectional material. Since these sequences are to be expected in any Raramuri discourse,[5] it makes sense to include them as a third type of unit for the synthesizer. However, the language is not sufficiently studied and these sequences are not really known. Fortunately, diverse unsupervised methods for the discovery of morphemes exist that con be applied to a corpus in order to determine these suffix sequences.

Table 2. A few of the 98 function words of Raramuri

pronouns	**determiners**	**pospositions**	**adverbs**
nihé	ecí	yuwa	chabé
muhé	mí	hiti	sinibí
ecí	ná	jonsa	gará
tamuhé		okua	arigá
tumuhé		pacháami	wabé
yémi		mobá	wikabé

Once the units were identified, a native speaker recorded each one in a natural context. The resulting waveform was labelled and segmented in order to compile the database that is used for TTS synthesis.

3.1 Segmentation Methods

Many techniques for morphological segmentation exist.[6] Some interesting ones are minimal distance methods [14], bigram statistics [15], minimization of affix

[5] In essence, lexical items —specifically the root morphemes within them— are the carriers of discourse. They convey content information in action. Also, some morphological items, specifically modifiers, clitics and affixes, which are derivational and inflectional, carry the grammatical information that structures discourse. Hence, one might argue that the essence of language as a communication system —which is embodied in its repeatable patterns— resides in its structure or in the items which structure discourse, like affixes. Therefore, sequences of these can be taken as a promising unit for a synthesizer, while the roots of the words in which they appear can be build by means of diphones.

[6] There are several prominent approaches to word segmentation. The earliest one is due to Zellig Harris, who first examined corpus evidence for the automatic discovery

sets [16] and Bayesian statistics [17]. For the purposes of identifying from a corpus a set of suffix sequences of Raramuri, any of these methods can be applied. We used an economy-entropy based method which we have previously applied to Spanish [18]; Chuj (Mayan) [19], Czech [20] and Raramuri [21].

The approach proposed in this paper grades word substrings according to their likelihood of representing an affix or a valid sequence of affixes. The resulting candidates are gathered in a table for later evaluation by experts. In essence, two quantitative measurements are obtained for every possible segmentation of every word found in a corpus: Shannon's entropy [22] and a measure of sign economy [13] (which will be dealt with below). In short, the highest averaged values of these two measurements are good criteria to include word fragments as items in the table which will be called affix *catalog*, *i.e.* a list of affix candidates and their entropy and economy normalized measurements, ordered from most to least affixal.

Information Content High entropy measurements have been reported repeatedly as successful indicators of borders between bases and affixes [18, 19, 23–25, 20]. These measurements are relevant because shifts of amounts of information can be expected to correspond to the amounts of information that a reader or hearer is bound to obtain from a text or spoken discourse. Frequent word fragments contain less information than those occurring rarely. Hence, affixes must accompany those segments of a text which contain the highest amounts of information.

Information content of a set of word fragments is typically measured by applying Shannon's method.[7] In order to identify affix sequences, the task is to measure the entropy of the word fragments which occur concatenated to a suffix candidate: where there is an actual morphological border, the content of information of stems with respect to their accompanying suffix sequences exhibits a peak of entropy. Specifically, looking for peaks of information means taking each right-hand substring of each word of the sample, determining the probabilities of everything that precedes it, and applying to these Shannon's formula to obtain the entropy measurements to be compared.

Economy Principle The other important measure used to identify Raramuri suffix sequences is based on the principle of economy of signs. In essence, we

of morpheme boundaries for various languages, [11]. His approach was based on counting phonemes preceding and following a possible morphological boundary: the more variety of phonemes, the more likely a true morphological border occurs within a word. Later, Nikolaj Andreev designed in the sixties the first automatic method based on character string frequencies which applied to various languages. His work was oriented towards the discovery of whole inflectional paradigms and applied to Russian and several other languages, [12]; and that of [13] in the seventies for French and Spanish.

[7] Recall the formula $H = -\sum_{i=1}^{n} p_i \log_2 p_i$, where p_i stands for the relative frequency of word fragment i [22].

can expect certain signs to be more economical than others because they relate to other signs in an economical way. Specifically, affixes combine with bases to produce a number (virtually infinite) of lexical signs. Although affixes do not combine with any base, certain ones combine with many bases, others with only a few. Nevertheless, it makes sense to expect more economy where more combinatory possibilities exist. This refers to the syntagmatic dimension. The paradigmatic dimension can also be considered: as they attach to bases, affixes appear in complementary distribution in a corpus with respect to other affixes (*i.e.* they alternate in that position). If there is a relatively small set of alternating signs which adhere to a large set of unfrequent signs the relations between the former and the latter must be considered even more economical.

The economy of segmentations can be measured by comparing the following sets of word fragments from each word of the corpus. Given a suffix candidate, there are two groups of word fragments:

1. *companions* — strings beginning graphical words which are followed by the given suffix sequence candidate (syntagmatic relation).
2. *alternants* — strings ending graphical word which occur in complementary distribution with the suffix sequence candidate.

Formally, let $A_{i,j}$ be the set of *companions* occurring, according to a corpus, along with word segment $b_{i,j}$. Let $A^p_{i,j}$ be the subset of $A_{i,j}$ consisting of the word beginnings which are quantitative prefixes of the language in question. Let $B^s_{i,j}$ be the set of word endings which are, also according to the corpus, suffixes of the language and occur in complementary distribution (*alternants*) with the word fragment $b_{i,j}$. One way to estimate the economy of a segmentation is:

$$k^s_{i,j} = \frac{|A_{i,j}| - |A^p_{i,j}|}{|B^s_{i,j}|} \tag{1}$$

In this way, when an right-hand word fragment is given, a very large number of companions and a relatively small number of alternants yield a high economy value. Meanwhile, a small number of companions and a large one of alternants indicate a low economy measurement. In the latter case, the word fragment in question is not very likely to represent exactly an affix, nor a sequence of them.

3.2 Building a Catalog of Raramuri Suffixes

The process to identify suffix sequences basically takes the words of the word sample and determines the best segmentation for each one using to the two measurements discussed above. Each best segmentation represents a hypothesis postulating a base and a suffix sequence. Thus, the presumed suffix sequence (and the values associated with it) are fed into a structure called Catalog.

The methods described above complement each other in order to identify Raramuri suffix sequences. Specifically, the values obtained for a given word fragment are normalized and averaged. That is, we estimated the *suffixality* of each sequence by means of the arithmetic average of the relative values of entropy

and economy: $(\frac{h_i}{\max h} + \frac{k_i}{\max k}) * \frac{1}{2}$, where h_i stands for the entropy value associated to suffix candidate i; k_i represents the economy measurement associated to the same candidate; and $\max h$ returns the maximum quantity of h calculated for all suffixes (same idea for $\max k$).

As mentioned above, Raramuri is more an agglutinative language than a fusional one and word formation is mainly accomplished by means of suffixation. As could be expected, stems are followed by derivational suffixes, and these by inflectional ones. Since stems can be the result of other morphological processes, there might be morphemes to be discovered towards the beginning of words, but they are not necessarily affixal [21].

The corpus[8] corresponds to the Raramuri's variant from San Luis Majimachi, Bocoyna, Chihuahua. For today's corpora standards, this sample is a very small one, consisting of no more than 3,584 word-tokens and 934 word-types. Even though we cannot assume this sample's representativity of this variant, we proceeded to apply the method because it is robust for small corpora. Table 3 shows partial results of procedure.

Table 3. The 20 Most Affixal Raramuri Suffix Sequences

rank	suffix	frec.	squares	economy	entropy	affixality
1.	∼ma	35	1.00000	1.00000	0.88030	0.98050
2.	∼re	77	0.79960	0.81100	0.86060	0.82370
3.	∼sa	33	0.63640	0.93060	0.75590	0.77430
4.	∼ra	62	0.66130	0.64610	0.85080	0.71940
5.	∼si	28	0.75000	0.52570	0.83450	0.70340
6.	∼na	25	0.41140	0.72240	0.79840	0.64410
7.	∼go	4	0.21430	0.90650	0.64930	0.59000
8.	∼é	49	0.16620	0.43580	1.00000	0.53400
9.	∼ame	51	0.25210	0.30640	0.85910	0.47250
10.	∼gá	18	0.40480	0.37810	0.61360	0.46550
11.	∼ka	19	0.25560	0.28060	0.84130	0.45920
12.	∼á	67	0.13860	0.31330	0.91950	0.45710
13.	∼ré	11	0.16880	0.41020	0.73430	0.43780
14.	∼ga	50	0.18290	0.28340	0.80650	0.42430
15.	∼a	281	0.10520	0.18960	0.97250	0.42250
16.	∼ba	8	0.21430	0.30220	0.74000	0.41880
17.	∼ayá	8	0.21430	0.44320	0.57570	0.41110
18.	∼í	42	0.10200	0.26480	0.80540	0.39070
19.	∼či	39	0.10260	0.27510	0.74000	0.37260
20.	∼e	164	0.15240	0.29100	0.64290	0.36210

[8] Mainly texts collected by Patricio Parra.

Although Raramuri has only a few inflectional forms, the larger catalog exhibits more items containing inflectional material than were expected.[9] In fact, if inflectional suffixes are to be considered somehow more affixal than derivational ones, it should not be surprising to find the four most prominent Raramuri inflection affixes appear at the top of the table: ∼*ma*, ∼*re*, ∼*sa*, and ∼*si*, which mark tense, aspect and mode.

Using her own field work experience and taking into account the work of other experts, Alvarado determined the 35 most prominent nominal and verbal derivational suffixes for this language. 25 of these occurred within the first 100 catalog entries (a recall measure of 71% within this limit). The other entries are chains of suffixes (including sequences of derivational and inflectional items) and residual forms.[10] The 10 derivational suffixes which did not appear in the catalog are essentially verbal derivational forms, or modifiers of transitivity or some semantic characteristic of verbal forms. This might mean that the small sample used is more representative of nominal structures, rather than of verbal ones. These missing suffixes were added to the set of units to be processed for the synthesizer. Nevertheless, it is worth stressing that a significant part of the known Raramuri derivational system —essentially the nominal subsystem— was retrieved from a very small set of texts, which hardly constitutes a corpus of this language.

4 Stages of Text Processing for Unit Selection Synthesis

In general, the stages of text processing —which, as mentioned above, were achieved for the target language— are:

Transcription. It consists of a phonetic representation of the input text to be synthesized, keeping all punctuation and stress marks to preserve intonation clues; a new text (transcribed) will be created. It includes conversion of dates and numbers to a phonologic level.

Diphones division. The transcribed file is analyzed in order to extract its diphones, an output file is created with a list of them, organized from the most to least used. This program is sensitive to detect diphones according to their prosody and intonation. The extraction must be capable to identify the stressed syllable in each word according to the stress rules of the language (and assigns a stress mark to its related diaphone). It also takes into account which punctuation is adjacent to the last diphone to identify its intonation.

[9] This is certainly due to the fact that input texts are constituted by linguistic acts in the pragmatic act of narrating a story. Words appear therefore inflected. Obviously, using dictionary entries without inflection (lemma sets), rather than text in context, would be a much better way to obtain derivational items.

[10] The examination of residual items was especially difficult. Questions about lexicalized affixes (possibly fossilized items) and about the relationship between syllable structure and affix status emerged. These matters remain to be revised by Raramuri experts. Meanwhile, for evaluation purposes, entries with unexpected syllabic structure were not counted as acceptable suffixes nor valid sequences of them.

Most frequent words and affix sequences searching. The corpus used is the valuable source to consult the most common words in this language. Besides the diphones, the program must also be capable to identify these words.

Preliminaries of the recording session. Materials extracted from the corpus are compiled and arranged in contexts that the speaker could read in the recording session.

5 Processing

This module will be able to choose the set of units of the speech corpus, that better adjusts to a series of characteristics. The selection will be made so that it diminishes the total cost, sum of the unit costs and costs of concatenation between units. Equation 3 describes the difference between the target segment (u_i) and the candidate segment (t_i). Using equation 3 we get the concatenation cost, which reflects the smoothness of the concatenation between selected segments (u_{i-1}, u_i).

This module will be the one in charge to concatenate the different units that have been chosen by means of the selection algorithm. Consequently, it will be necessary to implement another module that obtains the phonetic marks of each file of wave from the corresponding curve. Computacional load will be reduced as posible for the developed programs. A strategy for the organization of the data base of units will be followed that allows to accelerate the searches.

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^u(t_i, u_i) \tag{2}$$

$$C^c(u_{i-1}u_i,) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) \tag{3}$$

Through the direct concatenation of units, we expect to get a good quality of synthesized speech because of the use of a large database and the definition of prosodic targets. Nevertheless, in order to increase synthesized speech quality, *i.e.* to make that the transitions between units are not perceivable, it will be necessary to make a processing on the result by means of algorithm TD-PSOLA.

This algorithm is used since it can be applied directly to the audio signal without the need for parametric extraction as is the case with LPC and other common algorithms used. For this algorithm to work we need to add a pitch-mark extraction phase to the database creation. This step is done offline so it carries no speed penalties during run-time.

For pitch-mark extraction we have used a dynamic programming based algorithm presented by Vladimir Goncharoff and Patrick Gries [26]. This algorithm was found to be very straightforward and highly reliable and gave out practically no extraction errors. An added bonus is that source code for the algorithm is distributed freely.

During run-time the speech signal is first divided into overlapping Hanning windowed pichmark-centered segments. The lengths of these windows must be larger than a pitch period and proportional to the pitch period. To achieve pitch modification these segments are aligned to the new position of the pitchmarks and the segments are then added together. A normalization values is also calculated from the Hanning window to eliminate energy modifications due to the overlapping [8].

It may also be necessary to duplicate or eliminate segments to maintain the duration of the signal at different pitchs or to accommodate time modification of the signal simultaneously to pitch modifications. To minimize discontinuities in the concatenation points we use this algorithm to modify the pitch of each segment, so as to make the pitch of both segments equal. The segments are then cropped so that their beginning and end correspond to a pitch mark and their magnitude is equalized [8].

Although this process is unable to eliminate all discontinuities these are greatly minimized, but at the cost of very little distortion compared to other OLA based systems with heavier processing and the processor load is also smaller to achieve acceptable speed in slower equipment

6 Closing Remarks

In this paper we have presented a method to develop a TTS synthesizer based on unit-selection for just about any language whose words are structured as a stem and a sequence of affixes, either derivational, inflectional or both. In this manner, building the utterances becomes a matter of selecting chains of function words, diaphones (to build stems) and suffix sequences to complete content words appearing in their contexts.

We expect this strategy to result in more natural speech than what could be expected using diphones alone. However, disadvantages remain in clearness, especially when comparing this simple system to top international systems. Nevertheless, this synthesizer uses a smaller amount of memory in comparison to those systems, which are tailored for some widely and very well known language like English or German.

Certainly, the system proposed can be improved in several ways, but the method can be readily applied to new languages in order to obtain relatively good synthesizers for them. Meanwhile, we are working to improve the clearness without naturalness degradation.

Acknowledgments. The work reported in this paper has been supported by a DGAPA UNAM grant for PAPIIT Project IN402008 *Glutinometría y variación dialectal.*

References

1. Campbell, N., Black, A.: Prosody and the Selection of Source Units for Concatenative Synthesis. In: Progress in Speech Synthesis. Springer Verlag (1995)
2. Hunt, A.J., Black, A.W.: Unit Selection in a Concatenative Speech Synthesis System using a large Speech Database. ATR Interpreting Telecommunications Research Labs
3. Zhao, Y., Liu, P., Li, Y., Chen, Y., Chu, M.: Measuring Target Cost in Unit Selection with kl-divergence Between Context-Dependent HMMs. Microsoft Research Asia, Beijing, China, 100080
4. Dutoit, T.: An Introduction to Text-to-Speech Synthesis. In: VoiceXML Review. Kluwer Academia Publishers, Netherlands (1997)
5. Huang, X., Acero, A., Hon, H.: Spoken Language Processing. Prentice Hall PTR (2001)
6. van Santen, J., Sproat, R., Olive, J., Hirschberg, J., eds.: Progress in Speech Synthesis. Springer (1997)
7. Black, A.W.: Speech Synthesis for Educational Technology. In: SLaTE Workshop on Speech and Language Technology in Education. (2007)
8. del Río, F., Herrera, A.: A Mexican Spanish Synthesis System Using a Pitch Syncrhonous Overlap Add. In: Proceedings of the IASTED International Conference on Signal and Image Processing. (2004)
9. Tatham, M., Morton, K.: Developments in Speech Synthesis. John Wiley, Chichester (2005)
10. Wightman, C.W., Syrdal, A.K., Stemmer, G., Conkie, A., Beutnagel, M.: Perceptually Based Automatic Prosody Labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-to-Speech Synthesis. In: ICSLP 2000. Volume II., Beijing (October 2000) 71–74
11. Harris, Z.S.: From Phoneme to Morpheme. Language **31**(2) (1955) 190–222
12. Cromm, O.: Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev. Abschluß des Ergänzungsstudiums Linguistische Datenverarbeitung, Frankfurt am Main (1996)
13. Kock, J.d., Bossaert, W.: The Morpheme. An Experiment in Quantitative and Computational Linguistics. Van Gorcum, Amsterdam, Madrid (1978)
14. Goldsmith, J.: Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics **27**(2) (2001) 153–198
15. Kageura, K.: Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences. Journal of Quantitative Linguistics **6** (1999) 149–166
16. Gelbukh, A., Alexandrov, M., Han, S.Y.: Detecting Inflection Patterns in Natural Language by Minimization of Morphological Model. In: Congreso Iberoamericano de Reconocimiento de Patrones, CIARP-2004. LNCS (2004)
17. Creutz, M., Lagus, K.: Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In: Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Finland (June 2005)
18. Medina-Urrea, A.: Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes. Journal of Quantitative Linguistics **7**(2) (2000) 97–114
19. Medina-Urrea, A., Buenrostro Díaz, E.C.: Características cuantitativas de la flexión verbal del chuj. Estudios de Lingüística Aplicada **38** (2003) 15–31

20. Medina-Urrea, A., Hlaváčová, J.: Automatic Recognition of Czech Derivational Prefixes. In: Proceedings of CICLing 2005. Volume 3406 of Lecture Notes in Computer Science. Springer, Berlin/Heidelberg/New York (2005) 189–197
21. Medina-Urrea, A., Alvarado-García, M.: Análisis cuantitativo y cualitativo de la derivación léxica en ralámuli. In: Primer Coloquio Leonardo Manrique, Mexico, Conaculta-INAH (September 2004)
22. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)
23. Hafer, M.A., Weiss, S.F.: Word Segmentation by Letter Successor Varieties. Information Storage and Retrieval **10** (1974) 371–385
24. Frakes, W.B.: Stemming Algorithms. In Frakes, W.B., Baeza-Yates, R., eds.: Information Retrieval, Data Structures and Algorithms. Prentice Hall, New Jersey (1992) 131–160
25. Oakes, M.P.: Statistics for Corpus Linguistics. Edinburgh University Press, Edinburgh (1998)
26. Goncharoff, V., Gries, P.: An algorithm for accurately marking pitch pulses in speech signals. In: International Conference Signal and Image Processing. (1998)