Published in Proceedings 14th Int. Cong. Phonet. Sciences, San Francisco, CA, August 1999.
For more recent work by Hertz and her collaborators, see http://www.novaspeech.com/publications.htm

LANGUAGE-UNIVERSAL AND LANGUAGE-SPECIFIC COMPONENTS IN THE MULTI-LANGUAGE ETI-ELOQUENCE TEXT-TO-SPEECH SYSTEM

Susan R. Hertz*†, Rebecca J. Younes*, and Nina Zinovieva*
**Eloquent Technology, Inc., Ithaca, NY*
†*Department of Linguistics, Cornell University, Ithaca, NY*

ABSTRACT

ETI-Eloquence is a multi-voice, multi-language, rule-based text-to-speech system based on the innovative Delta synthesis technology [4, 6, 8]. In the last three years, the system has been extended from General American English to nine new languages and dialects, including British English, Mexican and Castilian Spanish, Canadian and Parisian French, Mandarin Chinese, German, Italian, and Brazilian Portuguese. One of the goals of this work has been to develop a universal strategy for text-to-speech rule development and as large a common rule base as possible. This paper describes the result of this work, providing an overview of the language-universal and language-specific components underlying our current rule system.

1. INTRODUCTION

Over the last sixteen years, Hertz and her associates have developed an innovative technology for multi-language and multi-dialect text-to-speech synthesis. At the heart of this technology is the Delta System rule development tool, which contains a powerful programming language [4] and an interactive development environment [8] specifically designed for formulating and testing rules that operate on a special type of multi-tiered utterance representation called a *delta*. As illustrated below, a delta consists of *streams* of coordinated units, such as phrases, words, syllables, phonemes, and acoustic values. Unlike the more restrictive SRS system from which it evolved [3], the Delta System is a highly flexible tool with which linguists can develop rules based on a wide range of phonological and phonetic models. The focus of this paper is on the particular models and strategies for text-to-speech synthesis that we have developed with Delta and incorporated into the multi-language ETI-Eloquence rule system.

For each language, the ETI-Eloquence rules are divided into two main components: the *text module* and the *speech module*. On the basis of the input text, the text module generates the delta, which depicts the abstract linguistic structure of the utterance. The speech module uses this representation to generate appropriately timed acoustic parameter values for a formant synthesizer, which produces the speech waveform.

The sections that follow describe each of the two modules in turn, illustrating how the rules build a delta through a combination of language-universal and language-specific components. While the paper focuses on the division of rules into these two types of components, it should be noted that, in the case of languages for which we generate more than one dialect, our rule system is also divided into large dialect-universal and smaller dialect-specific components in both the text and speech modules [7]; a discussion of these, however, lies outside the domain of this paper.

2. THE TEXT MODULE

For each language, the text module performs two main functions: *text normalization* and ***text parsing.*** The text normalization rules read the input text into the input stream of the delta utterance representation, determine the location of sentence boundaries, and generate a string of fully spelled out words, as shown in Figure 1 for the English sentence *It rained 5 in. so we went in*.[1] The text parsing rules operate on the output of the text normalization rules to build the linguistic structure of the utterance, including phrases, words, morphemes (if relevant for the language), syllables, and phones, as illustrated in the delta fragment in Figure 2 for the sample sentence. (The symbol # in the figure represents a pause.)

The units in each stream are separated by vertical bars called *sync marks*, which are used to coordinate units across streams. All vertical bars in the same column represent the same sync mark. For example, the sync marks immediately to the left and

```
sentence:  |sent                                                            |
input:     |I|t|' '|r|a|i|n|e|d|' '|5      |' '|i |n   |  .|' '|s|o|' '|w|e|' '|w|e|n|t|' '|i|n|.|
text:      |i|t|' '|r|a|i|n|e|d|' '|f|i|v|e|' '|i|n|c|h|e|s|' '|s|o|' '|w|e|' '|w|e|n|t|' '|i|n|.|
```

Figure 1. Delta fragment after text normalization

```
sentence:  |sent
inton_phr: |phrase                                        |   |phrase
text:      |i|t |' '|r|a|i|n|e|d|' '|f|i  |v|e|' '|i|n|c |h|e|s|' '|s|o |' '|w|e |' '|w|e|n|t|
word:      |wrd |   |wrd        |   |wrd      |   |wrd          |   |wrd |   |wrd |   |wrd    | . . .
morph:     |root|   |root   |suf|   |root     |   |root    |suf|   |root|   |root|   |root   |
syllable:  |syl |   |syl        |   |syl      |   |syl|syl      |   |syl |   |syl |   |syl    |
phone:     |I|t |   |r|e  |n|d  |   |f|a|y|v  |   |I|n|t|š |ɨ|z|#  |s|o |   |w|i |   |w|E|n|t|
```

Figure 2. Delta fragment after text parsing

right of the text stream characters `es` of *inches* in the delta in Figure 2 are also defined in the morph and phone streams, aligning the letters with a suffix unit in the morph stream and the phones `ɨz` in the phone stream. The sync mark to the left of `es`, while defined in the morph stream, is not defined in the syllable stream, since the morph boundary does not coincide with a syllable boundary.

In addition to the information shown in Figure 2, the units in each stream contain further information representing relevant linguistic properties. For example, for all the languages, each intonational phrase unit contains information about tonal characteristics of the phrase (e.g., high vs. low boundary tone); each word unit contains information about its grammatical category and degree of prominence in the phrase; each syllable unit contains information about degree of lexical stress and pitch accent type, when relevant (using a Pierrehumbert-type of analysis [1, 10]); and each phone contains information about its place and manner of articulation. Since the ETI-Eloquence system provides annotations with which users can specify voice characteristics on a word by word basis, each word unit in the delta also contains information about its voice characteristics, including degree of breathiness, vocal tract size, pitch range, and overall pitch level.

The main purpose of the text module is to produce the linguistic information needed for the derivation of acoustic values by the speech module. However, some of the information generated by the text module is not used directly by the speech module, but is needed for subsequent analysis within the text module itself. For example, grammatical category information is generally not referred to directly within the speech module, but is used for all languages in the prediction of phrase boundaries and intonational properties of words and phrases. In several of the languages, grammatical categories are used to disambiguate homographs (e.g., English *to present* vs. *the present*); in French, grammatical categories are also used in the prediction of liaison. Similarly, morphological information is used primarily in the prediction of phonemes and lexical stress (cf. English *naked* vs. *baked*, *prejudice* vs. *prejudge*).

The construction of a delta for an utterance involves both language-universal and language-specific strategies. For example, to produce the normalized text stream sequence, the text normalization module contains an outer universal procedure that parses the input stream into tokens and sends these tokens to language-specific, context-sensitive realization rules according to the token type (e.g. digit sequence, potential acronym, potential abbreviation, punctuation, and so on).

In our sample sentence, for example, language-universal rules recognize the periods as potential abbreviation or end of sentence markers. Language-specific rules determine that the first period delimits an abbreviation. In the case of the second period, since no language-specific interpretation rules apply, the universal rules assign the default end of sentence interpretation and insert the sentence unit into the delta. This close interplay between language-universal and language-specific rules is further illustrated by the treatment of expressions such as *1975-1987*. Language-universal rules determine that this type of expression is likely to represent a sequence of years, and communicate this information to the language-specific rules that generate the actual realization of the numbers and the hyphen.

After the text normalization rules have inserted fully spelled-out words into the text stream, the text parsing rules perform the linguistic analysis of the utterance, filling in the linguistic streams to provide the speech module with the information necessary to predict the acoustic values. Like the text normalization rules, the text parsing rules incorporate a number of universal strategies that control the overall processing. For example, language-specific phrase prediction rules determine the location of potential phrase breaks, using grammatical categories and other language-specific information, while universal procedures select the actual phrase breaks from among the candidates using a variety of criteria, including a language-specific minimum number of words per phrase.

During the course of rule development, new language-universal generalizations may emerge. For example, we have noted that there are particular grammatical structures that trigger intonational phrase breaks in all the languages. As we observe such universals, we refine the language-universal and language-specific modules accordingly. (Conversely, should we note during development of rules for a new language that certain universals are not applicable, we can prevent the rules in question from applying to this language using a general mechanism we have developed for “tagging” rules for particular sets of languages.)

3. THE SPEECH MODULE

The speech module uses the linguistics information produced by the text module to determine perceptually-relevant synthesizer parameter values and durations for the utterance. Depending on the desired voice quality, the parameter values are further modified by a set of language-universal voice filters to produce selected voice characteristics (male, female, child, breathy, rough, high-pitched, etc.). The final parameter values are sent to a Klatt-style formant synthesizer [9], which produces the final speech waveform.

The rules of the speech module are based on the phone-and-transition model of segmentation developed by Hertz [5, 6, 7], which makes possible the straightforward expression of both language-universal and language-specific generalizations concerning acoustic patterns, as discussed below. Roughly, phones represent those portions of the second formant pattern in spectrograms that can be attributed to the articulation of a particular speech segment, while transitions represent those portions that result from the movement of the articulators from one phone to another.

To produce the phone and transition structure, the speech rules for all languages first insert transition units with accompanying durations between each pair of adjacent phones. It has become apparent through our multi-language work that many of the transition durations can be determined by language-universal rules that are sensitive to the place and manner of articulation of the phones in question. We plan to factor out the universal transition duration rules into a language-universal component as our work progresses.

After the transitions have been inserted, the phones and transitions are grouped into larger timing units. Of particular significance is the “acoustic nucleus,” which consists of the syllable nucleus plus any voiced transitions into and out of the first and last nuclear phones, as illustrated in the delta fragment in Figure 3 for the word *five* of our sample sentence.[2]

```
phone:        |f  |  |a |  |y |  |v |
transition:       |tr|  |tr|  |tr|  |
acoustic_nuc:        |nucleus    |  |
ms:           |110|30|84|75|10|20|60|
```

Figure 3. Delta fragment after transition and nucleus insertion

In our system for English, the nucleus contains the vowel of the syllable plus any following tautosyllabic sonorants.[3] As shown in Figure 2, the nucleus of *five* is realized in our system with two phones (a and y), while the nucleus of *rained* is realized as a single phone (e), even though the phone e is diphthongized in most contexts in the dialect of American English the rules produce. The determination of whether a particular "gliding vowel" should be treated as one phone or two is made empirically, on the basis of timing patterns and other phonetic criteria [2, 7].

For each language, language-specific rules assign a total duration to the acoustic nucleus based on its composition and context. A language-universal procedure subtracts the transition and non-vowel phone durations within the acoustic nucleus from the total nucleus duration, and assigns the remaining duration to the vowel, thereby capturing a general trading relationship between the vowel and non-vowel durations of the acoustic nucleus [2].

The nucleus-based phone-and-transition structure provides a language-universal template for the positioning of synthesizer parameter values. In all languages, for example, formant values are generally aligned by the rules at the edges of each phone; voicing amplitude values are positioned at the beginning and end of each acoustic nucleus; aspiration for [h] and aspirated stops (if relevant) is aligned with a transition [5]; and a stop burst is positioned at the rightmost edge of the stop phone with which it is associated [5, 8].[4]

The delta fragment in Figure 4 shows the alignment of the voicing amplitude (AV), the frication amplitude (AF), and the second formant values (F2) that the rules would produce for our default voice (adult male) for the word *five* of our sample utterance.[5] A voicing amplitude of 0 dB (i.e., no voicing) and a frication amplitude of 55 dB are aligned with the 110 ms long phone f. At the beginning of the following 30 ms formant transition between the phones f and a (which is also the beginning of the acoustic nucleus), frication amplitude is turned off and voicing amplitude of 52 dB is turned on.

The formant values associated with 0 ms durations represent non-steady-state targets (i.e., inflection points) that help shape the overall formant pattern. Values between adjacent sync marks in a stream are interpolated over the specified duration when the final synthesizer parameter values are produced. For example, during the transition from the phone f to the phone a the second
formant moves from the value of 1000 Hz at the end of the fricative to 1200 Hz at the beginning of the vowel. During the vowel, the second formant moves from 1200 Hz to 1400 Hz over a period of 84 ms.

The rules that generate formant values assign language-specific values to each phone for the default voice initially produced by the rules, and then modify these values according to the phone’s context. For example, the first value of the phone f is raised due to the alveolar segment that precedes it in the sample sentence. Similarly, the different formant values at the edges of the phone a result from the operation of the coarticulation rules. It has become clear through our rule development that a large subset of the coarticulation rules that make such context-sensitive modifications are common across languages, and can be factored out into a separate language-universal component. Like other parameter values, each formant value generated by the rules may be further modified by the relevant language-universal voice filters to produce the desired voice characteristics for the word.

Like amplitude and formant values in each language, F0 values are positioned in relation to phone and nucleus units (though they are not necessarily aligned at their edges). For each syllable that has been assigned a pitch accent, or tone, by the text module, a language-specific procedure determines the fundamental frequency values and associated positions needed to realize the tone in question. The F0 rules are sensitive to well-known factors such as pitch range and degree of prominence of the word containing the syllable, as well as to the pitch properties of the particular voice being generated. In addition to fundamental frequency values for the pitch accents, the rules also generate values for phrase and boundary tones, in accordance with a Pierrehumbert-type model [1, 10].

Figure 5 shows the F0 values that are produced by the rules for our default voice for the words *five inches* in the sample sentence. The high tone associated with the stressed syllable of each of these words is realized by two F0 values—113 Hz and 122 Hz for *five* and 104 Hz and 121 Hz for *inches*. The rules for English always position the second F0 value for a tone a percentage of the way through the acoustic nucleus of the syllable containing the tone (the "accented syllable"). Other values, which are used to shape the F0 contour for the tone, are positioned in preceding or following phones, which are not necessarily in the same syllable or word as the accented syllable. In addition to the F0 values for the pitch accents, the delta fragment contains two additional values (92 Hz and 85 Hz), which realize the low phrase and boundary tones associated with the intonational phrase.[6]

From acoustic information of the sorts shown in Figures 4 and 5, the system performs the necessary interpolations and generates 5 ms frames of fully-specified values that are sent to the formant synthesizer, which produces the speech waveform.

```
phone:        |f             |  |a             |  |y   |  |v   |
transition:                  |tr|              |tr|    |tr|    |
acoustic_nuc:                |nucleus                  |
AV:           |0             |52                       |30  |
AF:           |55            |0                        |55  |
F2:           |1300|   |1000|  |1200|   |1400|  |1775|  |1300|
ms:           |0   |110|0   |30|0   |84 |0   |75|10  |15|60  |
```

Figure 4. Delta fragment after insertion of selected acoustic values

```
phone: |f        |  |a |         |y |  |v       |  |I         |  |n | |t | |š |  |ɨ       |  |z  |
F0:    |  |113|           |122|          |104|          |121|                        |92|    |85 |
ms:    |55|0  |55|30|84|46|0  |29|10|15|6|0  |54|40|40  |0  |5|20|55|1|35|1|90|40|20|0 |9|18|110|
```

Figure 5. Delta fragment after insertion of F0 values

4. CONCLUSION

This paper has presented the overall methodology underlying the multi-language ETI-Eloquence text-to-speech system, focusing on the common strategy used in the synthesis of the nine languages/dialects for which we have developed rules. The universal elements stem from the fact that the same kinds of linguistic information are relevant for the generation of acoustic values in all the languages, and similar kinds of analyses are involved in deriving this information from an input text. The large number of universals within the speech module reflect common physiological factors underlying the production and perception of human speech.

The uniform strategy underlying the synthesis of each of the languages has both practical and theoretical advantages. On the practical side, the universal components, coupled with the powerful Delta System tools, result in extremely rapid development of rules for different languages. For example, in a matter of just over a year, a staff of nine linguists at Eloquent Technology, Inc. developed complete text-to-speech systems for five languages. On the theoretical side, our language-universal framework provides a powerful basis for investigating hypotheses about the phenomena that we observe in particular languages, and straightforwardly capturing the generalizations that emerge within the language-specific and language-universal components of our rule system.

ACKNOWLEDGEMENTS

Development of the multi-language text-to-speech system was funded in part by DARPA under Cooperative Agreement #F30602-96-2-0196 administered by Rome Laboratory, by Contract #RW97077002 from the U.S. Department of Education, and by IBM.

NOTES

1. In order to present our overall strategy without going into details that are beyond the scope of the paper, the sample deltas here and elsewhere in the paper differ in selected details from the actual deltas generated by the ETI-Eloquence rules at the time of writing.

2. In the case of adjacent sonorants belonging to different nuclei (e.g., the [i] and [æ] of *piano*), the transition between the sonorants is either assigned entirely to one of the two nuclei or divided between them, depending on a number of phonetic criteria.

3. The status of nasals in relation to the acoustic nucleus is unclear. For some purposes, nasals seem to function as part of the acoustic nucleus, and for other purposes they do not. We are still investigating this question.

4. These generalizations abstract away from spectrographic details that we have determined to be perceptually irrelevant.

5. See [2] for an alternate system for relating acoustic values to phonological units based on autosegmental representations.

6. In general, a sync mark is positioned in the millisecond (ms) stream at every point that an acoustic value starts or ends. For clarity, however, when selected streams are displayed, only the relevant sync marks are shown, and duration units are collapsed accordingly. The differences in the ms stream between Figures 4 and 5 reflect this notational convention.

REFERENCES

[1] Beckman, M. E. and J. B. Pierrehumbert. 1986. Intonational structure in English and Japanese. *Phonology Yearbook,* 3, 255-310.

[2] Clements, G. N. and S. R. Hertz. 1995. An integrated approach to phonology and phonetics. In Durand, J. and B. Laks (eds.), *Current trends in phonology; models and methods.* CNRS, Paris X and University of Salford Publications.

[3] Hertz, S. R. 1982. From text to speech with SRS. *Journal of the Acoustical Society of America,* 72, 1155-1170.

[4] Hertz, S. R. 1990. The Delta programming language: an integrated approach to non-linear phonology, phonetics, and speech synthesis. In Kingston, J. and M. Beckman (ed.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge University Press.

[5] Hertz, S. R. 1991. Streams, phones, and transitions: toward a phonological and phonetic model of formant timing. *Journal of Phonetics* 19, *Special Issue on Speech Synthesis and Phonetics*, edited by R. Carlson.

[6] Hertz, S. R. 1999. The ETI-Eloquence text-to-speech system. White paper available at http://www.eloq.com.

[7] Hertz, S. R. and M. Huffman. 1992. A nucleus-based timing model applied to multi-dialect speech synthesis by rule. *Proceedings of the International Conference on Spoken Language Processing,* 2, 322-325.

[8] Hertz, S. R. and L. Zsiga. 1995. The Delta System with Syllt: Increased capabilities for teaching and research in phonetics, *Proceedings ICPhS* 95, 2, 322-325.

[9] Klatt, D. H. and L. Klatt. 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America,* 87, 820-857.

[10] Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT.