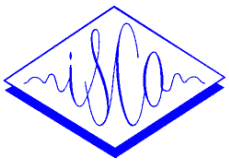

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A NUCLEUS-BASED TIMING MODEL APPLIED TO MULTI-DIALECT SPEECH SYNTHESIS BY RULE

Susan R. Hertz and Marie K. Huffman

Eloquent Technology, Inc.
24 Highgate Circle, Ithaca, N.Y. 14850
and
Dept. of Modern Languages and Linguistics, Cornell University

ABSTRACT

This paper presents a new timing model for rule-based speech synthesis which underlies the rules we are developing for five American English dialects, and which we are beginning to extend to other languages. The model leads to extremely efficient development of high-quality rules for different dialects and languages, and, more generally, provides new insights into the nature of speech. The paper presents the basic tenets of the model, showing how it leads to generalizations about speech patterns within and across dialects (and languages) that cannot be captured in more conventional models. The introduction discusses the nature of utterance representations structured in accordance with our model, using an example from General American English. The second section illustrates our application of the model to rule-based synthesis of American English dialects, focussing in particular on the straightforward and accurate rules for formant timing made possible by the model. The third section discusses the directions our work is currently taking, including development of a novel, multi-dialect relational database for extracting cross-dialect and intra-dialect generalizations, and experiments aimed at determining when observed variability in timing patterns within and across dialects is perceptually important.

I. INTRODUCTION

In conventional models for formant-based synthesis by rule, speech is treated as a linear sequence of phoneme-sized segments (e.g., [1][2]). Formant target values are positioned relative to these segments, with formant transitions overlapping adjacent segments or incorporated into individual segments. Hertz [3]-[5] has discussed in some detail how the utterance representations used in these models lead to unnecessarily complex rules and how they obscure cross-dialect and cross-language universals. This paper presents a new timing model for rule-based synthesis in the acoustic domain that has grown from our own experience developing rules for different languages within a conventional framework [1][6]. Within this model, we have been able to capture cross-language, cross-dialect, and intra-dialect generalizations, particularly concerning formant timing, that are virtually impossible to express with conventional models.

One of the fundamental ways in which our model differs from conventional models is in the basic segmentation of speech patterns that is assumed. Rather than treating speech as a sequence of abutted, phoneme-sized segments, we view speech as a sequence of separate *phone* and *transition* units, organized into higher-level linguistic units. Figure 1, for example, shows a spectrogram of the word *tied* (as in *Say tied for me*), with a conventional segmentation above it, and the segmentation underlying our rules below it.

In the conventional segmentation, which is used, for example, for the rules developed by Klatt [2], the diphthong `ay` is treated as a single unit that includes the aspirated formant transition from `t` into the diphthong, the transition between the two components of the diphthong (`a` and `y`), and the voiced transition from the diphthong into `d`. Hertz [3]-[5] has shown how this segmentation and similar segmentations obscure many generalizations about speech, such as the fact that aspiration and glottalization tend to align with formant transitions, that phones (steady-state portions) in some contexts lengthen independently of transitions, that diphthongs behave durationally like other tautosyllabic vowel-sonorant sequences, and that formant transitions in certain contexts have similar durations across languages.

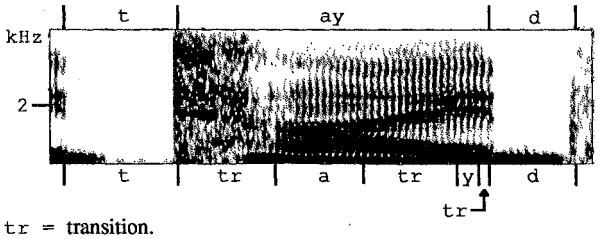

Fig. 1. Conventional (top) vs. our (bottom) segmentation of *tied*.

With a segmentation into separate phones and transitions, in contrast, we can straightforwardly account for these and other generalizations about speech timing. In addition, we can give all phones across all languages and dialects identical underlying acoustic structures—e.g., formant targets positioned precisely at the edges of the phones, as illustrated below. An additional advantage of independent phones and transitions is that they lead to clearly prescribed guidelines for segmenting spectrograms, which in turn leads to mutually-consistent, integratable rules across languages and dialects.

A linear string of phone and transition units alone, however, would not suffice to account for the speech patterns observed in spectrograms. Instead, we must group these units into higher-level linguistic units, such as words and syllables. A particularly important unit, not recognized in other models, is the *acoustic nucleus*, which serves a basic organizational role within the durational system of all the languages we have investigated [5].

Consider, for example, the following partial representation (simplified somewhat for purposes of illustration) that our rules would build for the General American English (GA) word *tied*, as uttered in *Say tied for me:*

```
word:      |word                                    |
syll:      |stress1                                 |
diaphon:   |t    |    |ay              |    |d     |
phone:     |t    |    |a     |    |y   |    |d     |
F2:        |1800 |    |1350  |    |1900|    |1800  |
asp_amp:   |0    |70  |0                            |
voic_amp:  |0         |60                    |0     |
acst_nuc:                |nuc                  |
trans:     |     |tr  |      |tr  |      |tr  |      |
ms:        |95   |70  |55    |90  |20    |15  |60    |
```

This representation is called a *delta*, because it consists of interconnected *streams* like a river delta. The vertical bars, called *sync marks*, coordinate the units in the various streams. Among others, the non-numeric streams include `word`, `syll` (syllable), `diaphon` (diaphoneme), and `phone`. The `word` stream delimits words, while the `syll` stream delimits syllables and represents their degree of stress. The `diaphon` stream contains underlying "phonemes" common to all American English dialects, while the `phone` stream contains the phones particular to the dialect being synthesized [3]. For example, in GA the diaphoneme `ay` is realized as a diphthong, represented by the two phones `a` and `y`, as shown above, while in many Southern and Black dialects this diaphoneme would be realized by a single monophthongal phone in many contexts [3]. The diaphonemes are generated from text, with a very accurate set of rules we have developed over several years; these rules are not discussed further in this paper. In addition to those

shown above, deltas in our model have other non-numeric streams, including an intonational phrase stream that describes the tonal pattern of the utterance. Units in the non-numeric streams also have associated features, not shown in the delta fragment above. For example, phone units have features such as consonant, stop, vowel, etc.

Each unit in the phone stream is synchronized with second formant values in the F2 stream that characterize the phone, and the intervening tr (transition) tokens in the trans stream represent the movement from one formant target to the next. The phone t, for example, has a second formant target value of 1800 Hz, which is held constant for 95 ms, as specified in the ms stream. Similarly, the phone a has a second formant value of 1350 Hz, held for 55 ms. A transition lasting 70 ms separates the targets for t and a, with the formant values during the transition computed via interpolation.

The asp_amp (aspiration amplitude) stream represents the aspiration pattern. The 70 ms transition between t and a has 70 dB of aspiration, representing the fact that the stop aspiration overlaps this transition [3]-[5]; all other stretches of the delta have 0 dB of aspiration (i.e., no aspiration).

The voic_amp (voicing amplitude) stream represents the voicing pattern. Voicing (60 dB) starts at the beginning of the phone a and ends at the end of the transition into d.

In addition to streams for the second formant, aspiration amplitude, and voicing amplitude patterns, a complete delta for synthesis would contain a variety of other acoustic streams, including streams for the other formants, nasalization, fundamental frequency, and so on. While all of the values in the acoustic streams in the sample delta fragment above are aligned precisely with the edges of higher-level linguistic units, some acoustic values, such as fundamental frequency and amplitude targets, are sometimes positioned partway through such units. For example, [4] illustrates the early cessation of voicing partway through the transition into a syllable-final voiceless stop.

The acst_nuc stream represents the acoustic nucleus, which consists of the stretch of the delta containing the vowel of the syllable, any non-nasal tautosyllabic sonorants, and any voiced portions of transitions on either side. For *tied*, the acoustic nucleus consists of the phone a, the transition from a to y, the phone y, and the transition from y to d. The initial transition from t to a is not voiced, so it is not part of the acoustic nucleus. In died, in contrast, the initial transition would be voiced, so it would be part of the nucleus. Acoustic nuclei in English can have up to three phones. For example, the nucleus of *wild* contains the three phones a, y, and l.

Our research has shown the acoustic nucleus to be an important duration unit in that speakers tend to give the acoustic nucleus as a whole a particular duration, depending on its context [5]. The duration of the acoustic nucleus is distributed among the component phones and transitions of the nucleus in a principled fashion, as discussed in Section II. In addition to being an important duration unit, the acoustic nucleus also plays an important role in determining amplitude and fundamental frequency patterns. The acoustic nucleus is also important for distinguishing American English dialects in that the most salient perceptual differences among dialects seem to lie in the formant patterns within the nucleus.

II. THE RULES

At the time of writing, we have developed a comprehensive set of rules for GA founded on our nucleus-based phone-and-transition model. In a number of pilot studies (including a global study of five American English dialects and a more detailed study of a Black English dialect), we have verified that the model will apply equally well to other American English dialects. We are now beginning to develop comprehensive rules for four dialects of American English other than GA, one from each of the following families: New England, New York, Southern, and Black. The rules for each of these dialects will be integrated with the rules for GA into a single multi-dialect rule set, containing dialect-universal and dialect-specific components, as discussed below.

The multi-dialect rules are being developed with the Delta System, a sophisticated, linguistically-oriented software system for building and manipulating deltas [7]-[9]. The rules are expressed in the Delta programming language [8]-[9], and are tested and refined using DeltaTools, an interactive environment for tracing Delta programs and experimenting with rules [9]. While we are using the Delta System to implement the rules, the model underlying the rules is not in any way built into the system; Delta can be used to write rules structured in accordance with a wide range of models. Delta programs are compiled into a portable format for incorporation into end-user products.

The rules operate on text, building a multi-stream delta of linguistic and acoustic information, such as that shown for *tied* above. On the basis of the acoustic streams, values are generated for a Klatt synthesizer [11]. (It is important to note, however, that the synthesizer is not an integral part of the model, which could just as well be used to generate values for other formant synthesizers.)

Our GA duration rules have been based on an extensive study over many years of a single speaker of GA, although we have observed similar durational behavior for many other speakers of the dialect. The primary methodology used in developing the rules is cyclical formulation and testing of hypotheses, alternating between analysis and synthesis. In the analysis phase of each cycle, we study natural speech data, mainly by examining and taking measurements from spectrograms, from which we develop hypotheses about the factors accounting for the observed patterns. (Spectrograms are made with the Kay real-time DSP Sona-Graph (model 5500) and Entropic System's Waves/ESPS speech analysis software running on a Sun Workstation.)

In the synthesis phase, we use the Delta System to build deltas that embody the hypotheses formed during the analysis phase. Depending on the situation, we either formulate the hypotheses as rules within our Delta program, or test them through interactive experimentation using DeltaTools. In either case, we evaluate the resulting speech. Among other things, this evaluation includes listening to the speech back to back with natural speech, and visually comparing spectrograms of the synthetic and natural speech. We also administer periodic intelligibility tests, including tests using semantically anomalous sentences and isolated syllables. For the isolated syllables, we use the well-known Modified Rhyme Test [11] and more phonetically-balanced tests of our own. We plan to begin administering formal naturalness tests soon. Visually, the generated formant patterns, particularly those in acoustic nuclei, match the patterns of our model speaker very well.

In the last several Modified Rhyme Tests we have administered, using an open response format, the intelligibility has generally been above 80%, with the scores improving steadily on successive tests. The results compare favorably with the 84% reported in [12] for DECtalk, the best of the eight text-to-speech systems tested in that study, especially when one considers that the test of the eight systems was administered in a carefully controlled setting over headphones, while ours were administered over a loudspeaker, without controlling for room noise.

2.1 Overview

The rules have two main parts: text-to-phone and phone-to-speech. The text-to-phone rules generate the underlying, multi-stream linguistic description of the utterance from the input text, while the phone-to-speech rules generate the acoustic patterns from the linguistic description. This section focusses on the phone-to-speech rules, particularly on how they use the acoustic nucleus and its constituent phone and transition units to produce formant patterns. A discussion of the text-to-phone rules is beyond the scope of this paper.

The phone-to-speech rules are continually being refined and restructured as we examine more data, particularly for new dialects and languages, and their present organization differs somewhat from that presented in earlier papers. Since the Delta System results in clear and maintainable rules, we can efficiently refine and reorganize the rules as we learn more about speech.

The rules for formant patterns are divided into *dialect-universal* rules, which we believe will apply to all or most dialects, and *dialect-specific* rules, which generate values for particular dialects. With a few exceptions, the dialect-universal rules generate the formant patterns for stretches of the delta outside of the acoustic nucleus, while the dialect-specific rules generate the patterns inside the nucleus. Even in the dialect-specific rules, however, a number of generalizations can be made across dialects, particularly in the *form* of the duration rules, as discussed below. In the actual rule set, language-universal rules are also applied. Among other things, the language-universal rules generate an underlying abstract

acoustic representation common to all languages, as discussed in [3]. The language-universal component, however, is not critical to an understanding of our basic algorithm for generating formant patterns for English dialects, and is not discussed further in this paper.

2.2 Dialect-Universal Rules

The dialect-universal rules are applied first. The following delta fragment shows the values these rules would produce for *tied*, as uttered in *Say tied for me*:

```
phone:     |t          |   |a |   |y |    |d          |
F2:        |1800|   |1800|   |  |   |  |    |1800|   |1800|
acst_nuc:  |                 |nuc          |
trans:     |               |tr |   |tr|   |tr |             |
ms:        |0    |95|0     |70 |   |90|   |15 |0    |60|0    |
```

(Here and elsewhere, only the relevant streams are shown.) Each phone outside of the nucleus (here, t and d) is given a durationless second formant value at the very beginning and end of the phone, a language-universal structure made possible by independent transitions. The phone duration connects the two targets. In addition, transitions are assigned durations. We expect many transition durations to be not only dialect-universal, but language-universal as well. Some transition durations, however, such as that between phones in the acoustic nucleus (e.g., between a and y above) may prove to be dialect-specific, in which case the rules for them will be moved to the dialect-specific rule component.

2.3 Dialect-Specific Duration Rules

Next, the dialect-specific duration rules apply. These rules begin by assigning to each acoustic nucleus the duration it would have in the frame *Say* [b__t] *for me*, a context in which preliminary studies revealed similar durations for acoustic nuclei in different dialects. For example, a nucleus consisting of the phones a y is assigned a duration of 145 ms, while a two-phone nucleus ending in r is assigned 160 ms. Subsequent rules modify the starting duration according to the actual context, as discussed below.

After assigning a total duration to the nucleus, the rules assign durations to the non-vowel phones in the nucleus (recall that transition durations have already been assigned). For example, they assign to the y of *tied* a duration of 10 ms, as shown below. The total nucleus duration (145 in this case) is stored as a feature of the nucleus, not shown in the display.

```
phone:     |a  |   |y  |   |
F2:        |   |   |   |   |
acst_nuc:  |nuc          |
trans:     |   |tr |   |tr |
ms:        |   |90 |10 |15 |
```

Next, the duration of the vowel is determined by subtracting the total duration of the transitions and non-vowels in the nucleus from the total nucleus duration. That is, the vowel is pliable, receiving whatever duration will yield the correct total nucleus duration—in this case, 30 ms:

```
phone:     |a  |   |y  |   |
F2:        |   |   |   |   |
acst_nuc:  |nuc          |
trans:     |   |tr |   |tr |
ms:        |30 |90 |10 |15 |
```

Thus, the rules correctly handle the trading relationship between vowels and transitions in that the longer the transitions in the acoustic nucleus are, the shorter the vowel will be, as discussed in some detail in [5]. Note that the phone a of *tied* will correctly be given a longer duration than the a of died, since the nucleus of *tied* does not include the transition into the a, while the nucleus of died does, as discussed above. (The a of *tied* is shortened slightly by a subsequent rule that shortens phones after aspiration, as discussed below, but it remains longer than the a of *died*.)

Next, the rules modify the durations of nuclei in specific contexts. For example, a nucleus before a tautosyllabic voiced consonant (with the exception of nuclei before a tautosyllabic nasal-voiceless stop sequence) is lengthened to 1.5 times its current duration, or to a specified maximum duration, whichever is less. The maximum duration depends on the phone structure of the nucleus. For example, a three-phone nucleus is given a maximum duration of 250 ms, while a nucleus consisting of a vowel plus a glide (ay, oy, or aw) is given a maximum of 200 ms. In the case of *tied*, the rules would lengthen the nucleus from 145 ms to 200 ms.

When nuclei are lengthened, most of the lengthening occurs in the phones, rather than in the transitions, so the rules lengthen phones and transitions separately. The lengthening of a nucleus before a voiced consonant is carried out exclusively in the phones: First, any non-vowel phones are lengthened to a duration that depends on the particular phone; then the vowel phone is given whatever duration is needed to yield the correct total nucleus duration, in a fashion similar to that discussed above for determining the vowel's starting duration. The following delta fragment shows the new phone durations resulting for the nucleus of *tied*:

```
phone:     |a  |   |y  |   |
F2:        |   |   |   |   |
acst_nuc:  |nuc          |
trans:     |   |tr |   |tr |
ms:        |75 |90 |20 |15 |
```

One additional duration rule would apply, shortening the phone a to 55 ms due to the preceding aspiration [4]. In general, the factors that determine the final duration of an acoustic nucleus include the segmental context of the nucleus, its degree of stress, whether it is in a function or content word, and its position in the word and phrase that contain it.

Our preliminary work on other American English dialects suggests that durations of acoustic nuclei in different dialects are modified in the same kinds of contexts, with the dialects differing primarily in degrees of lengthening and shortening and in maximum durations. Thus, we expect to be able to "parameterize" many of the current duration rules for GA, to produce a dialect-universal set of duration adjustment rules for American English, with dialect-specific variables for the lengthening and shortening percentages and the maximum durations.

2.4 Dialect-Specific Formant Rules

After assigning durations, the dialect-specific rules assign formant values to the phones in the nucleus. Like the dialect-universal rules, the dialect-specific rules generally position these values at the beginning and end of each phone (although there are a few exceptions). The rules include a separate procedure for each phone that assigns the appropriate formant values to the phone via a set of context-sensitive rules. For *tied*, the procedures for a and y assign the values shown below:

```
phone:     |a                |   |y              |   |
F2:        |1350|      |1350|   |1900|   |1900|
acst_nuc:  |nuc                                      |
trans:     |                 |tr |                 |tr|
ms:        |0    |55    |0    |90 |0    |20|0     |15|
```

Each formant target is assigned a duration of 0 ms, even though in natural speech, targets often exhibit durations. Our perceptual tests have shown these target durations not to be important perceptually, so we have abstracted away from such details in the rules.

While each phone in the nucleus of *tied* receives the same values at the beginning and end of the phone, phones often have different values at each edge. Some phones intrinsically have different values at the beginning and end, while others have different values only in specific contexts. After all formant and other acoustic values have been generated, identical adjacent values are collapsed, producing a delta like that shown in Section I above.

III. CURRENT DIRECTIONS

To expedite our rule development for new American English dialects, we have undertaken two complementary projects: (1) development of a multi-dialect (and multi-language) relational database structured in accordance with our model, and (2) experiments designed to evaluate the perceptual relevance of variations observed in formant timing among speakers within and across dialects.

3.1 Multi-Dialect Database

The database is organized into three logical levels: utterance, syllable, and acoustic, each represented by one or more relational tables. Among other things, the utterance table includes overall

properties of each utterance, such as the text (spelling), phonetic context (e.g., frame, isolation), rate of speech, speaker, language, and dialect. The syllable table contains information for each syllable, such as degree of lexical stress and the context of the syllable (whether it is in a word with focus, whether it is in an unstressed word, whether it is phrase-final, etc.).

The acoustic tables are the heart of the database, containing phones and transitions as basic units, and associated acoustic data. Each phone is paired with its following transition, and numbered sequentially within its syllable. For each phone, the acoustic tables contain the formant target values at the edges of the phone, the duration of each formant target, the duration of the stretch between the formant targets, and the location and duration of any periods of glottalization, noise, aspiration, voicing, and nasalization. For the transition that follows the phone (connecting it with the next phone), the acoustic table includes the duration of the transition, and the location and duration of glottalization, noise, aspiration, voicing, and nasalization. Linked to the acoustic tables via the phone names is a table that gives the inherent features of a phone (e.g., its place and manner of articulation).

The database is currently implemented with the Paradox relational database system from Borland International, running on an IBM PC networked to our Sun Workstations, where speech is analyzed in preparation for entry into the database. We plan to port the database to our Sun computers, which will better accommodate the large amount of multi-dialect and multi-language data to be entered. We are using Entropic System's Waves/ESPS speech analysis software to segment spectrograms into phones and transitions, to mark other relevant information, and to automatically extract durations and formant values. We enter the information obtained in this fashion into the database with a customized computer program we have written.

We have verified the utility of the database in a number of pilot projects involving Black English, General American English, and German data. By querying the database, we can easily obtain durations and acoustic values in specified contexts within or across the dialects and languages; from the data extracted, we can derive rules. In addition to being an indispensable tool for synthesis rule development, the database should prove extremely valuable for speech recognition work as well.

3.2 Perceptual Experiments

While the database just described will provide a wealth of cross-dialect and intra-dialect information from which to extract generalizations, it will not provide information about when variability among speakers of different dialects is perceptually insignificant, and, hence, does not need to be modelled for purposes of synthesis. Even speakers of the same dialect exhibit consistent, but perceptually insignificant, timing differences, especially in how the total nucleus duration is distributed among the nucleus components. For example, in two speakers of General American English, we have observed consistently different distributions of the total nucleus duration in nuclei consisting of the phones `I l`, as in *build*. One speaker has a consistently longer `l` and shorter `I`, with the total nucleus duration quite similar for both speakers. In nuclei containing the phones `e l`, on the other hand, the speakers exhibit much more similar phone durations.

In order to determine when observed timing differences are perceptually significant, we have designed a series of experiments. Using the Delta System, we are generating groups of synthetic stimuli containing the same total nucleus duration, but different internal phone and transition durations. In informal tests of nuclei containing the phones `I l` and `e l`, we have found that in nuclei with `I l`, trading off phone durations has very little perceptual effect, while in those containing `e l`, the `e` must have a certain minimum duration in order to sound like a tense vowel. This result may explain why we found variability in `I l`, but not `e l`, for the two speakers. Speakers of the same dialect are obviously much more likely to exhibit different timing patterns if each pattern produces the same result perceptually. Continued experimentation of the sort described here should give us insights into the perceptual phenomena that constrain articulation. With these insights, we will be better able to interpret variability among speakers, and separate dialect-specific from speaker-specific rules.

IV. FINAL REMARKS

This paper has presented a nucleus-based timing model that underlies the synthesis rules we are developing for different dialects and languages, focussing on its application to American English. The model leads to simpler and more accurate rules for the timing of acoustic parameters, to similarly-structured acoustic representations for all languages and dialects, and to a division of rules into language-universal, dialect-universal, and dialect-specific components. In general, the model is part of a more comprehensive technology for efficient development of high-quality synthesis rules. Besides the model, this technology includes the Delta System and the database. Jointly, the model, the Delta System, and the database enable faster and more economical development of high-quality synthesis rules than formerly possible, within a framework that allows us to learn more about speech across languages and dialects.

Acknowledgments

We thank Allard Jongman, Aaron Kaplan, and Timothy Weber for their helpful comments on drafts of this paper. Our multi-dialect work has been funded in part with funds from the U.S. Dept. of Education under Contracts RS89071002 and RS90087003 with Eloquent Technology, Inc. (ETI), and from Grant 1 R43 DC00758-01 to ETI. Related multi-language work has been supported with funds from the New York State Science and Technology Fundation, under grant RDG 89174 to Cornell University (for collaborative work with ETI.). The content of this publication does not necessarily reflect the views or policies of these agencies.

References

[1] S. R. Hertz. "From Text-to-Speech with SRS." *J. Acoust. Soc. Am.* 72, pp. 1155-1170, 1982.

[2] D. H. Klatt. "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence." *J. Acoust. Soc. Am.* 59, pp. 1208-1221, 1976.

[3] S. R. Hertz. "A Modular Approach to Multi-Dialect and Multi-Language Speech Synthesis using the Delta System." *Proceedings of the Workshop on Speech Synthesis*, European Speech Communication Association, pp. 225-228, 1990.

[4] S. R. Hertz. "Streams, Phones, and Transitions: Toward a Phonological and Phonetic Model of Formant Timing." *Journal of Phonetics* 19, pp. 91-109, 1991.

[5] S. R. Hertz. "The Timing of Phones and Transitions: Toward a Nucleus-Based Timing Model of English Duration." *Working Papers of the Cornell Phonetics Lab.* 7, pp. 135-149, 1992.

[6] S. R. Hertz and M. Beckman. "A Look at the SRS Synthesis Rules for Japanese." *ICASSP*, pp. 1336-1339, 1983.

[7] S. R. Hertz. "Delta: Flexible Solutions to Tough Problems in Speech Synthesis by Rule," The Official Proceedings of Speech Tech 88, Media Dimensions Inc., N.Y, 1988.

[8] S. R. Hertz. "The Delta Programming Language: an Integrated Approach to Non-Linear Phonology, Phonetics, and Speech Synthesis." in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (ed. by J. Kingston and M. Beckman), Cambridge University Press, pp. 215-257, 1990.

[9] R. Charif, S. R. Hertz, and T. Weber. *The Delta System, Version 2 User's Manual*, Eloquent Technology, Inc., 1992.

[10] D. H. Klatt and L. Klatt. "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers," *J. Acoust. Soc. Am.* 87, pp. 820-857, 1990.

[11] A. S. House, C. E. Williams, M. H. Hecker, and K. D. Kryter. "Articulation-testing methods: consonantal differentiation with a closed-response set," *J. Acoust. Soc. Am.* 37, pp. 158-166, 1965.

[12] J. S. Logan, B. G. Greene, and D. B. Pisoni. "Segmental Intelligibility of Synthetic Speech Produced by Rule," *J. Acoust. Soc. Am.* 86, pp. 566-581, 1989.