Synthesis of Connected Speech

John Coleman
Oxford University Phonetics Laboratory

1. Introduction

In this paper I will report on recent extensions to the nonconcatenative speech synthesis architecture employed in the YorkTalk and IPOX systems (Coleman 1992a, 1994, Local 1992, Dirksen and Coleman 1994). These systems have not to date been noted for verbosity: our primary goals have been to achieve the most natural-sounding speech that we can, as well as solving the technical challenges presented by using nonsegmental phonological representations in speech synthesis.

However, some work directed towards the goal of continuous speech has already been reported. Local (1992) presented a phonological analysis of assimilation in place of articulation at word boundaries which, when employed in YorkTalk, generated some convincing examples of assimilated forms in combination. Coleman (1994) gave a synopsis of our approach to the synthesis of arbitrarily long polysyllabic words, a task which includes 1) syllabic, prosodic (metrical) and morphological parsing; 2) phonetic interpretation of syllable parts, syllables, and metrical feet; 3) smooth connection of metrical feet. Here I will advance this line of research two small steps further, and report on 1) synthesis of vowel reduction (schwa formation) and syllabic sonorants, and 2) synthesis of metrical feet across word boundaries. I shall start by briefly reviewing the standard architecture for synthesis-by-rule.

2. The standard model of connected speech synthesis

The standard synthesis-by-rule architecture (e.g. MITalk; Allen et al. 1987) is segmental and derivational. That is, the units employed in phonological, phonetic and even parametric representations are short consonant- and vowel-sized segments of speech. The segments at different levels (phonological, phonetic and parametric, possibly also orthographic) are related to one another by application of an ordered sequence of rewrite rules. Computationally, systems of this kind consist of a cascade of finite-state transducers, where each transducer computes the operation of one rule. As an idealization, we regard the standard text-to-speech architecture as a transparent implementation of SPE-style phonology.

For example, consider the synthesis of the words "photograph", "photographer" and "photographic". Morphological analysis (e.g. MITalk DECOMP, primarily an affix-stripping program) at the orthographic level segments these orthographic strings into "photograph", "photograph+er" and "photograph+ic" respectively. The graphemes of the orthographic stem "photograph" may be mapped onto systematic phonemes by a set of SPE-style grapheme-to-phoneme rules to yield e.g.:

```
        Graphemes:      ph  o  t  o  g  r  a  ph        er      i  c
        Phonemes:        f  O  t  O  g  r  a  f         @       i  k
```

The phonemic stem /fOtOgraf/, with three full vowels, the first two of which are tense, is a single underlying abstract lexical representation that is invariant for all three forms "photograph", "photographer" and "photographic". It thus has the merit of generality, but it is not surface-true, as none of these three forms has two tense vowels followed by a third full vowel. In all three surface forms, at least one of the vowels is reduced. Consequently, a set of structure-building rules (e.g. stress assignment) and structure-changing rules (e.g. vowel shift, diphthongization, vowel reduction and stress movement rules) is applied to map the invariant phonemic stem into three different phonetic forms: ['f@ut@,graf], [f@'togr@f@] and [,f@ut@'grafik]. Finally, some connected speech processes such as flapping, vowel elision and others may be modelled by the application of further, postlexical rules, yielding surface phonetic representations such as ['f@uR@graf] or [f:'togr@f@]. Notice that the combination of mappings from orthography to surface phonetics preserves the scale and the order (roughly) of the units throughout the derivation: in other words, the mapping is a homomorphism. This preservation of scale and order is even manifest at the parametric level, where each surface phonetic unit is mapped into a vector of synthesis parameter targets. Numerical interpolation between these segmental targets is used to model the continuous transitions observable in real speech:

```
Phones          f   @ u R   @  g  r  a  f

                |   | | |   |  |  |  |  |

Parameters  P1  - \ _ / - \ _ _ / - \ - -
(schematic) P2  _ / - - \ _ / - \ _ _ / -
            ...
```

3. Nonsegmental synthesis

In the nonsegmental approach to synthesis exemplified by YorkTalk, IPOX and other systems of a similar kind, the mapping from phonemes to synthesizer parameters is computed not by a cascade of transducers but by a computational approach called syntax-directed translation (see e.g. Holub 1990: 183-192). This method has two separate parts: 1) PARSING of a syntactic structure (in this case, morphological, metrical and syllabic structure), and 2) INTERPRETATION of that structural analysis. This typically means traversing the analysis graph and mapping parts of the analysis graph onto parts of the output description. In the most typical use of this technique, compilation of high-level computer languages, the parsing stage analyses the structure of the programmer's high-level input, and the interpretation stage translates small fragments of the analysed high-level code into small fragments of the computer's machine code. In non-segmental synthesis systems, interpretation rules relate small pieces of phonological structure (e.g. prosodic constituent structure and distinctive features) onto synthesizer parameters directly, without any intermediate levels of representation. The structures computed in the parsing stage are not obviously homomorphic with the input. The fragments generated in interpretation correspond to (parts of) phonological constituents of all sizes, not just consonants and vowels. These parametric fragments are combined with one another not just or in particular using concatenation and interpolation (as in the standard architecture) - though these operation are employed sometimes - but more especially they are combined by overlapping them in various ways.

4. Overlapping of speech constituents

The overlapping nature of speech constituents has been noted by many phoneticians, including O"hman (1966), Fowler (1980), Liberman (1970) and Browman and Goldstein (1986). In speech synthesis, overlapping combination has been described at length by Browman and Goldstein (1986) and Coleman (1992b). As overlapping phonetic interpretation in non-segmental synthesis has been described at length elsewhere (Coleman 1992b, 1994), I shall only summarize some key examples here. The main relevance of phonetic overlap in this paper is that it is one of the principal means by which synthetic speech can be made to sound more connected or fluent, and less disjointed. In addition, the particular pattern of overlaps employed in YorkTalk and IPOX is critical to our models of a number of connected speech processes that in the standard generative approach are independent rules with no connection to one another.

First, following O"hman (1966), Perkell (1969), Fowler (1980) and Griffen (1985), we regard syllable-initial and final consonants as being overlaid on syllable nuclei, rather than concatenated to them, as in segmental synthesis. This organization assists in numerically modelling coarticulation, because consonants are overlaid on the vowels with which they are coarticulated. In addition, I shall show below that compressing a syllable affects the duration of the vowel primarily, and leaves the consonants unaffected. This effect is employed in our treatment of vowel reduction in faster syllables.

Second, within consonant clusters, obstruents are overlaid on sonorants. As an example of the consequences of this pattern, in the cluster /tr/, the /t/ takes on the labiality of the /r/ upon which it is overlaid. In the overlapping model, this simply means that the labiality of the the /r/ is not over-written or obscured by the /t/:

```
+---------------------------+
|       labiality of /r/    |
+----------+----------------+
|   /t/    |
+----------+
```

Third, in diphthongs, the targets of offglides are sensitive to the quality of the peak, so that, for example, the phonetic height of [i] is somewhat lower in [ai] than in [ei].

Fourth, in synthesis of connected sequences of syllables, the constituents at the margins of syllables overlap to different degrees, depending on their phonological identity.

5. Synthesis of metrical feet

5.1 Syllable juncture

Although we make extensive use of overlap as a way of joining phonological units together, we still need to be able to join some units end-to-end, in a concatenative fashion. Joining utterances together is one such case. However, if we simply concatenate two syllables in IPOX, e.g. /sit/ and /ti:/, we will not succeed in generating a fluent disyllabic word /sitti:/ "city". The final consonant of /sit/ in isolation is slightly aspirated, and is followed by silence. The initial consonant of /ti:/ is fairly strongly aspirated. Concatenation will result in a prounciation [sit’ thi:], containing two stop bursts, two periods of aspiration and a short pause. Although that pronunciation would not be correct for connected speech, we want to allow it as a possibility which will be employed in the case where /sit/ is at the end of one utterance, and /ti:/ begins a new utterance, with a deliberate pause by the speaker between the two utterances. This is the first type of syllable juncture, which we might call concatenative or pause juncture.

In order to make the two syllables cohere into a fluent disyllable, we begin the stop closure of the initial /t/ of /ti:/ before the closure of the final /t/ of /sit/ has been released. In this way, the burst, aspiration and final pause of /sit/ are all overwritten by the start of the second syllable. Depending on the degree of overlap between the two closures, we can generate a range of intervocalic stop closure durations by varying the degree of overlap between the two closure portions. The range is from hardly any closure at all, when the release of the /t/ of /ti:/ begins almost as soon as the closing phase of the final /t/ of /sit/, through closure durations typical of single intervocalic consonants, up to closure durations approaching twice as long as isolated syllable-initial consonants. The degree of overlap can be controlled by varying the time between the release time of the first stop and the release time of the second stop, labelled X and Y respectively in the diagram below. The three regions in the overlap continuum that I have just listed correspond to three well-defined types of consonant duration: 1) very short, 2) normal, and 3) long. Very short closure is our model of intervocalic "flapping". Long closure is our model of gemination, e.g. the distinction in English between the intervocalic consonant portions of "wholly" ([houlli] in most varieties of English) vs. "holy" ([houli]), "fish shop" [fiSSop] vs. "bishop" [biS@p], "that tin" [Dattin] vs. "satin" [satin], etc. In addition, pause juncture can be modelled by overlap rather than concatenation, simply by timing release time Y sufficiently long after release time X, making a fourth region on the syllable overlap continuum. The timing of various events in these four cases can be illustrated as follows.

Type 1. Y is before X: very short closure ("flapping")

```
               End of first syllable

 Vowel  Closing       Closure      Burst & Asp  Pause
-------|-------|-------------------|----------|------
                                   X

                          Y
                ----------|----------|------------------------------
                 Closure   Burst & Asp         Vowel

                                  Second syllable
```

Type 2. Y is approximately equal to X: normal length intervocalic closure

```
               End of first syllable

 Vowel  Closing       Closure      Burst & Asp  Pause
-------|-------|-------------------|----------|------
                                   X

                                   Y
                         ----------|----------|------------------------------
                          Closure   Burst & Asp         Vowel

                                  Second syllable
```

Type 3. Y is after X: long closure ("gemination")

```
               End of first syllable

 Vowel  Closing       Closure      Burst & Asp  Pause
-------|-------|-------------------|----------|------
                                   X

                                               Y
                                 --------------|----------|--------------------
                                      Closure   Burst & Asp         Vowel

                                  Second syllable
```

Type 4. Y is much later than X: pause juncture

```
               End of first syllable

 Vowel  Closing       Closure      Burst & Asp  Pause
-------|-------|-------------------|----------|------
                                   X

                                                              Y
                                                    ----------|----------|-----
                                                     Closure   Burst & Asp Vowel


                                                    Second syllable
```

The different degree of overlap in each of these four cases is in part determined by the phonological identity of the objects at the syllable margins. For example, the timing of stops is obviously different from the timing of fricatives or approximants. Also, juncture types 1 and 2 are only applicable when the syllable-final and syllable-initial consonants are phonologically identical, whereas juncture types 3 and 4 may be used whether the syllable-final and syllable-initial consonants are the same or different. But primarily, the choice of juncture type is governed by the linguistic context of the two syllables in question. As has often been observed, type 1 ("flapping") mainly occurs word-internally between a stressed and an unstressed vowel (see e.g. Kenyon 1924 [1961]:126-7). Type 2 occurs word-internally as well as at word boundaries where a word ending in a vowel meets a word beginning with a single consonant, or where a word-final consonant is joined to a word-initial vowel. We model these as containing an identical syllable-final and syllable-initial consonant, with the closure of a single consonant, so that the vowel-to-consonant and consonant-to-vowel transitions can be generated locally to each of the two syllables. Type 3 ("gemination") occurs especially in compounds (e.g. "nut-tree"), phrases (e.g. "that case"), and certain within-word morphological environments, especially with Germanic suffixes (e.g. "wholly", "evenness").

5.2 Joining syllables at their roots

Conjoining syllables is not simply a question of the temporal coordination of events at the end of one syllable and the beginning of the next. In addition, we must model holistic properties of prosodic units larger than syllables, as well as some properties that are specific to syllables in combination. As an example of the former, we shall briefly discuss the generation of F0 contours for metrical feet. As an example of the latter, we shall discuss the relative duration (compression) of syllables in combination and isolation. In the next section, we discuss a number of vowel and consonant quality differences between syllables in isolation and combination, arising naturally from different degrees of compression.

The overall F0 contour of a metrical foot in English synthesis in YorkTalk and IPOX is a simple linear high-to-low fall (sometimes transcribed H* L). The high peak is the phonetic interpretation of the strong first syllable of the foot. The low trough may be viewed as the phonetic interpretation of the end of the foot, perhaps contrastively cueing a following foot-initial high toned syllable, if one is present. The H* L pattern is found irrespectively of the number of syllables in the foot, in declarative utterances. (In interrogative utterances, and in other dialects, the pattern s w => L H may be found instead. We shall not examine such a case here.) See Pierrehumbert (1980) figs. 1.1A, 1.16, 1.22, 2.5, 2.34, 2.40C et al. for examples of the H* L intonation contour on words of two syllables and more. In addition to the overall pitch fall computed at the foot level, at lower levels of prosodic structure, individual consonants and vowels contribute small details to the complete F0 specification (e.g. so-called consonantal perturbations), similar to the F0 algorithm briefly discussed in Pierrehumbert and Beckman (1988: 176-7, fig 7.1d). In our architecture, those modifications are overlaid on the simple, foot-level linear falling contour, because phonetic interpretation is computed top-down, so that large-scale holistic details may be augmented by small-scale refinements at lower levels of prosodic structure.

5.3 Vowel quality reduction and syllabic coda sonorants

One of the phonetic properties computed for each syllable node is a variable called "compression", with arbitrary units representing the duration ratio of the syllable in a particular context relative to its isolation duration. (The compression variable is employed with other types of constituent too, but we shall limit this discussion to syllables.) Because of the overlapping nature of consonants, vowels and syllable margins, the notion of "syllable duration" can only be defined operationally. Intuitively, however, a compression of 0.5 yields a syllable token which is about half as long as its isolation pronunciation; 2.0 doubles the duration of the syllable, and so on. The compression value of a syllable must be carefully controlled, as we have determined experimentally that a large range of fine variations are found from one environment to another, principally conditioned by the number and nature of the other syllables in the foot. There may be other factors too, but I have only included factors which are local to each foot in the model.

Compression values of c. 0.75-0.85 give rise to reduced vowels, that is, shorter vowels with more central vowel qualities. The temporal reduction in the vowel arises from the facts that a) the duration of the vowel is set to the duration of the entire syllable, and b) consonants are overlaid on vowels. Overlaying fixed-duration consonants on a compressed vowel causes less of the inter-consonantal portion of the vowel to be audible. For example, compare the durational minimal pairs /t@:n/ (e.g. "turn") and /t@n/ (e.g. second syllable of "button"):

```
Compression:    1.0                      0.8
V Duration:     Longer                   Shorter

        +-----------------+      +-------------+
        |        @        |      |      @      |
        +-----+-------+---+      +---+----+---+
        |  t  |       | n |      | t  |    | n |
        +-----+       +---+      +---+    +---+
```

The quality of vowels in compressed syllables differs from isolation forms because of the fact that the consonant-to-vowel and vowel-to-consonant formant frequency and bandwidth transitions, being part of the consonant, are relatively fixed in duration. Compressing a syllable, therefore, brings the onset and coda consonants closer together. The vowel's acoustic parameters are only independently audible between the end of the onset transitions and the start of the coda transitions. If the consonants are very close together, the apparent vowel quality may reflect the influence of consonantal transitions. The vowel's acoustic targets may be unreached or overwritten by the consonantal transitions:

```
        +------------+
        |      @     |
        +-----++-----+
        | t  /  \  n |
        +---+    +---+
```

This effect may be regarded as an acoustic analogue of articulatory undershoot in shorter vowels. Only a modest degree of compression is needed to derive such rapid speech forms as [s:p>t] ("support") and [th:mAt@u] ("tomato"), in which the vowel of the first syllable has been completely eclipsed by the friction and aspiration phases, respectively, of the initial consonant.

Because sonorant consonants tend to have very long duration in coda position, compression of syllables containing a short vowel and a coda sonorant gives rise to apparent syllabic sonorants. To illustrate this, we generated two sets of tokens of the words /k^zin/ ("cousin") and /bot@l/ ("bottle"). (Similar demonstrations have been made for numerous other words containing coda sonorants.) Note that in /k^zin/ we specified the vowel of the second syllable as the lax high front vowel /i/, which gives rise to rather a hypercorrect spelling pronounciation in slow tokens. We varied the compression values of the two syllables of each word as follows:

```
        Word            Compression 1           Compression 2
```

cousin	1	0.75	0.8
	2	0.75	0.75
	3	0.75	0.7
	4	0.75	0.65
	5	0.75	0.6
bottle	1	0.88	0.95
	2	0.88	0.9
	3	0.88	0.825
	4	0.85	0.8
	5	0.8	0.8

Note that in "bottle" tokens 4 and 5 we reduced the compression value (i.e. reduced the duration) of the first syllable, not just the second syllable, in order to improve the rhythm.

The duration continua tabulated above gave rise to changes in the audible quantity and quality of the vowel of the second syllable, as well as to changes in the overall rhythm of the words. In addition, the secondary articulation of the [l] in the "bottle" tokens appeared to vary also. This is discussed below. First, I shall tabulate the subjective changes in nucleus quality:

Word		Nuclear quality	Comments
cousin	1	[i]	Hypercorrect
	2	[i"]	Slightly centralized
	3	[@+]	Very short, more central
	4	[(@)n]?	Fleeting [@] or syllabic [n]
	5	[n]	Syllabic [n]
bottle	1	[@+]	Clear [l]
	2	[@]	Not so clear [l]
	3	[@-]	Dark [l]
	4	[(@)l]	Fleeting [@] or syllabic [l]
	5	[l]	Syllabic [l]

I should emphasize that the changes in quality of the vowel and coda sonorant reported here are not merely perceptual illusions. The synthesizer parameter files show minor variations arising from the interaction of invariant consonant and vowel target parameters under different temporal conditions. I have already mentioned that the more central qualities of the more compressed syllables arise from an acoustic analogue of articulatory undershoot. The fact that the coda laterals appear to be darker in more compressed syllables was a little surprising, though it may be relevant to note Sproat and Fujimura's (1993:300) finding that the time interval between maximal closure of the tongue tip and maximal lowering or retraction of the tongue body is significantly longer in dark [l]'s than in clear [l]'s. If this is the case in natural speech, it seems reasonable to expect that in synthetic [l] tokens differing only or primarily temporally, apparently longer [l]'s may be perceived as darker than apparently shorter [l]'s, even if they are actually durationally and spectrally similar. In the present experiment, the [l]'s in "bottle" tokens 4 and 5 are apparently longer than in tokens 1, 2 and 3 because the preceding vowels are shorter, i.e. the [l]'s in 4 and 5 take up a greater proportion of the second syllable duration than in tokens 1 to 3.

6. Analysis and synthesis of metrical feet spanning word boundaries

As well as "connections" between syllables and sub-syllabic units, the phrase "connected speech" usually refers to connected sequences of words. Our work in this area is still very preliminary. Nevertheless, I would like to report on an important part of inter-word juncture, namely the analysis and synthesis of metrical feet that cross word boundaries. This is not greatly different from within-word metrical feet, so much of the means for generating sequences of words is already in place. I shall take as a working example the sentence "this is an adaptable system".

6.1 Analysis

The main problem posed by this sentence is a severe mismatch between morphosyntactic and prosodic structures. This is illustrated in the two representations below. The first is obtained by parsing orthographic input using simple IPOX grammars for English syntactic and morphological structure, in which phrase structure rules are annotated with metrical structure. The second is obtained by parsing phoneme-based input using a prosodic grammar for English. (Internal structure of syllables is not shown. Ambisyllabic consonants are written out twice, though the Oxford/YorkTalk system computes which consonants or clusters are ambisyllabic.)

Morphosyntactic structure: [this [is [an [[ad apt able] system]]]]

Prosodic structure: [[Dis siz z@n n@d] [dapt t@b b@l] [sist st@m]]

Although the two structures are radically different, they are systematically related, and a general solution would need to take both into account. In generative phonology, this relation is usually described in procedural terms: unstressed function words and stray initial syllables initially remain "unparsed", and are "adjoined" at a later stage to the preceding foot. In declarative theory, such a solution is not available, so the two structures are produced in parallel.

In the current example, the structure on the right is arrived at by requiring that heavy syllables such as /dapt/ and /sist/ only appear as the head of a foot, whereas light syllables are weak nodes of a foot (except phrase-initially). Such a restriction however is warranted only in the Latinate part of the lexicon (cf. Coleman 1994) and should not be generalized to phrase-level prosodic structure. Nevertheless, we can propose the following account of trans-word metrical parsing. First, trans-word metrical structure respects within-word metrical structure. As monosyllables, "this", "is", and "an" only have prosodic structure up to the level of the syllable. We cannot determine how they are grouped into feet at the word level, since they have no within-word sister syllables. "System" is a disyllabic foot at the lexical level: this information is respected at the sentence level.

"Adaptable" contains an initial stray syllable /@d/ and a complete metrical foot /dapt t@b b@l/. In some previous accounts (e.g. Pierrehumbert and Beckman 1988:148), initial stray syllables are not footed but are attached to the word node. I shall depart from that treatment in order to avoid restructuring when words are joined together. I shall treat the initial light syllable /@d/ as the final syllable of the trans-word foot /Dis siz z@n n@d/. In order to construct this trans-word foot, it suffices to note that neither "is" nor "an" is exceptionally accented. As light syllables, then, they are weak, just as they would be within a polysyllabic word. Trans-word feet must be unbounded, like within-word Germanic feet, if we are to derive the tetrasyllabic foot /Dis siz z@n n@d/.

The sentence consists of three trans-word feet, then: /Dis siz z@n n@d/, /dapt t@b b@l/ and /sist st@m/. We do not yet have a parser capable of analyzing utterances of more than one phrase or tone group, so we complete the analysis by grouping the three feet into a single phrase in which the final foot is the head.

6.2. Temporal interpretation

At the next stage, each node in the prosodic tree is assigned a duration, and for each branching node the amount of overlap between its siblings is determined (usually just looked up from a table). Subsyllabic constituents are assigned a duration on the basis of the feature structures associated with them in the analysis, as well as their prosodic environment. Durations of higher-level constituents are determined compositionally, taking overlap between constituents into account. Across syllables, the duration of a constituent is taken to be the sum of the durations of its siblings minus the amount of overlap between them. If no overlap (or a negative amount of overlap) is specified, syllables are simply concatenated. This type of juncture is appropriate between utterances and prosodic phrases, but not phrase-internally. For example, simply concatenating the syllables /sist/ and /t@m/ results in a pronunciation [sist' th@m], containing two stop bursts, two periods of aspiration and a short pause. In this case, it is necessary to make /st/ ambisyllabic, starting the second syllable somewhat before the end of the first syllable, such that the release of the unaspirated onset /t/ overwrites the release of the coda /t/, a type of juncture which is typical of ambisyllabic consonants.

Rhythmic properties of an utterance are accounted for by compressing the durations of syllables which appear in polysyllabic feet. For example, while a monosyllabic foot is assigned a compression factor of 1.0, the stressed syllable of a disyllabic foot might be assigned a compression of 0.9, and its weak sister one of 0.8. The rhythm of feet consisting of three or more syllables can be approximated by a compression pattern such as 0.85...0.65...0.75. However, actual values for syllable compressions also depend on the internal structure of syllables. For synthesis of "this is an adaptable system" we have found the syllable compression sequence [0.92, 0.8, 0.8, 0.775, 0.816, 0.7802, 0.55, 1.1, 0.77] to yield adequate results. The tabulation of compression values for various types of words is a focus of our ongoing research.

7. Conclusion

I have described a method of analyzing a sentence prosodically and computing phonetic parameters for the Klatt synthesizer from such an analysis. A number of phenomena that are unrelated in a more conventional system based on rewrite rules, such as coarticulation, unstressed vowel shortening, centralization of unstressed vowels, syllabic sonorant formation and elision of unstressed vowels before syllabic sonornants are modelled in YorkTalk and IPOX as natural concomitants of the all-prosodic view of phonological structure and phonetic interpretation. In the future we hope to improve the quality of our phonetic parameters, as well as addressing the prosody-syntax interface more thoroughly.

Acknowledgement

I would like to thank my collaborator Arthur Dirksen (Institute for Perception Research, Eindhoven, The Netherlands) for his implementation of IPOX, and assistance in preparing this paper and the audio recordings played at the workshop. Latter parts of this paper are not disjoint from Dirksen and Coleman (1994), which is shorter, but has better pictures.

References

Allen, J., M. S. Hunnicutt and D. Klatt (1987) From text to speech: the MITalk system. Cambridge: Cambridge University Press.

Browman, C. P. and L. M. Goldstein (1986) Towards an articulatory phonology. Phonology 3, 219-252.

Coleman, J. S. (1992a) "Synthesis-by-rule" without segments or rewrite-rules. In G. Bailly, C. Benoit and T. R. Sawallis, eds. Talking Machines: Theories, Models and Designs. Amsterdam: Elsevier. 43-60.

Coleman, J. S. (1992b) The phonetic interpretation of headed phonological structures containing overlapping constituents. Phonology 9.1. 1-44.

Coleman, J. S. (1994) Polysyllabic words in the YorkTalk synthesis system. In P. A. Keating, ed. Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III. Cambridge: Cambridge University Press. 293-324.

Dirksen, A. and J. S. Coleman (1994) All-prosodic synthesis architecture. Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, September 12-15, 1994.

Fowler, C. A. (1980) Coarticulation and theories of extrinsic timing. Journal of Phonetics 8. 113-133.

Griffen, T. D. (1985) Aspects of Dynamic Phonology. Amsterdam: Benjamins.

Holub, A. I. (1990) Compiler design in C. Englewood Cliffs, NJ: Prentice Hall International.

Kenyon, J. S. (1924) [1961] American Pronunciation. Ann Arbor, MI: George Wahr Publishing Co. Tenth edition.

Liberman, A. M. (1970) The Grammars of Speech and Language. Cognitive Psychology 1. 301-323.

Local, J. K. (1992) Modeling assimilation in nonsegmental, rule-free synthesis. In G. J. Docherty and D. R. Ladd, eds. Papers in Laboratory Phonology II: Gesture, Segment, Prosody. Cambridge University Press. 190-223.

O"hman, S. E. G. (1966) Coarticulation in CVC utterances: spectrographic measurements. JASA 39. 151-168.

Perkell, J. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. Cambridge, MA: MIT Press.

Pierrehumbert, J. B. (1980) [1987] The Phonology and Phonetics of English Intonation. Indiana University Linguistics Club.

Pierrehumbert, J. B. and M. E. Beckman (1988) Japanese Tone Structure. Cambridge, MA: MIT Press.