Proceedings of The Institute of Acoustics

Cosegmentation in the IBM Text-to-Speech System

J B Pickering

Speech Research, IBM UK Scientific Centre,
Athelstan House, St Clement Street, WINCHESTER, SO23 9DR

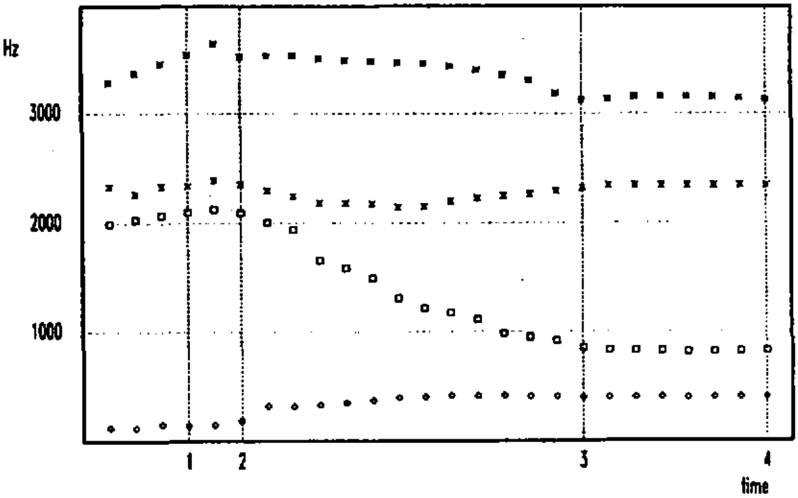
Introduction

This paper outlines some of the considerations leading to the cosegmentation rules in the IBM text-to-speech system. *Cosegmentation* is used to refer to two processes: (i) the interpolation of parameter values across segment boundaries (the generation of *transitions*), and (ii) the modification of temporally disparate values due to the effects of immediate segmental context (*coarticulation* proper). We shall concentrate on two aspects of the cosegmentation rules here: first the temporal organization of segment transitions, and second the context-dependent changes in formant parameters. The analysis begins with a range of /CV/ combinations, which forms the basis of the parameter library and of the cosegmentation rules. Tracking the parameters through the /CV/ demonstrates how parameters change across segment boundaries, which helps in modelling transitions, and necessarily includes modifications to parameter values due to immediate segmental context.

Data Collection and Analysis

A single male RP speaker with no previous phonetic training was recorded in a sound-proof room at the IBM UKSC reading lists of all phonotactically possible symmetrical /CVC/ syllables at a subjectively comfortable speed. To avoid the intrusion of some kind of 'list-intonation', items were included at the beginning and end of each section which were not subsequently analysed. The recordings were digitized at 10 kHz with a 4.5 anti-aliasing filter and then edited using the IBM PC-based digitization system [1]. Subsequent analyses were undertaken using the host-based signal processing facilities [2]. The digitized speech was then analysed every 10 msecs with LPC covariance techniques, usually across a 256-point analysis window. (For stop consonants, the release was analysed using a 100-point window.) Formant centre frequencies and bandwidths were extracted by root-solving, and formant amplitudes by automatically scaling the FT of the filter coefficients to a power spectrum. Misassigned peaks were corrected by eye. This procedure gave a set of formant parameter files for each of the /CVC/-syllables. The initial processing of these parameter files involved only the /CV-/ portions for each syllable. Four frames for each /CV-/ combination where isolated, representing < 1 > a notional consonant steady-state (or the release for the stop consonants), when the parameters were not changing appreciably, < 2 > the offset of the consonant, < 3 > the onset of the vowel, and < 4 > the vowel steady-state. Figure 1 shows the syllable [jɔ:], with the position of each of these four frames indicated. We shall consider the three first formant centre frequencies for these frames.

Figure 1: Formant tracks for the /CV-/ [jɔː]



Transition durations

The duration of the transition between consonant and vowel is of course perceptually significant. It is possible, for example, to shift the percept of the prevocalic consonant [b] to [w] by manipulating the length of the transition. We begin, therefore, by considering the durations of the transitions in our own RP corpus. The number of 10 msec frames between frames <2> and <3>, representing the transition proper for each /CV-/ combination, was measured as a gross estimate of the time taken to move from one segment to the next. This was taken as the transition duration. A two-way ANOVA (vowel×consonantal context) showed that consonantal context has a significant effect on the duration of the transition ($F[18,180] = 15.354$, $p < 0.01$), while vowel quality produces only a marginal effect ($F[10,180] = 2.151$, $p \leq 0.05$). If (phonologically) long and short vowels are treated separately, it turns out that long vowels are associated with slightly longer transition durations (61.3 msec) than the short vowels (52.5 msec, $F[1,9] = 18.404$, $p < 0.01$). The influence of specific consonantal contexts on transition durations is discussed in more detail below.

First consider whether *voicing* or *place of articulation* affect the duration of the transition. For example, would voiced stop consonants be associated with a shorter transition, since the voicing characteristics of the consonant and vowel are the same? Similarly, would the transition for labial consonants be shorter than for other places of articulation, since the tongue body is free to move during the articulation of the consonant into a position suitable for the following vowel? Three-way fixed effects ANOVAs (vowel×voicing characteristic×place of articulation) were run on the transition data for stop consonants and for fricatives independently. For both consonant types, none of the interactions was significant at the 0.05 level. Only the voicing characteristic produced a significant main effect ($F[1,20] = 19.726$ for the stop consonants and $F[1,20] = 19.747$ for the fricatives, $p \leq 0.01$ in both cases). For the stops, the mean transition duration for voiceless consonants was 44.8 msec, significantly longer than the mean duration for voiced consonants at 33.9 msec. For the fricatives, the reverse was true: the mean transition duration for the voiceless consonants was 47.6 msec, and for the voiced consonants 62.7 msec. In neither case was the place of articulation a significant factor ($F[2,20] = 0.015$ for the stops, and 0.814 for the fricatives). Further tests on the nasals [m,n] and the glides [w,j] also revealed place of articulation to be insignificant ($F[1,20] = 0.237$ for the nasals, and 1.586 for the glides).

Next, consider the *manner of articulation*. The average transition durations for each vowel across the voiceless stops and the average duration across the voiced fricatives where compared with the average durations for the glides and for the nasals, and the measured durations for [l,r]. Manner of articulation proved to be significant ($F[5,60] = 18.846$, $p < 0.01$). The voiceless stops are associated with the shortest average transition durations (44.8 msec) and the consonant [r] with the longest (105.5 msec).

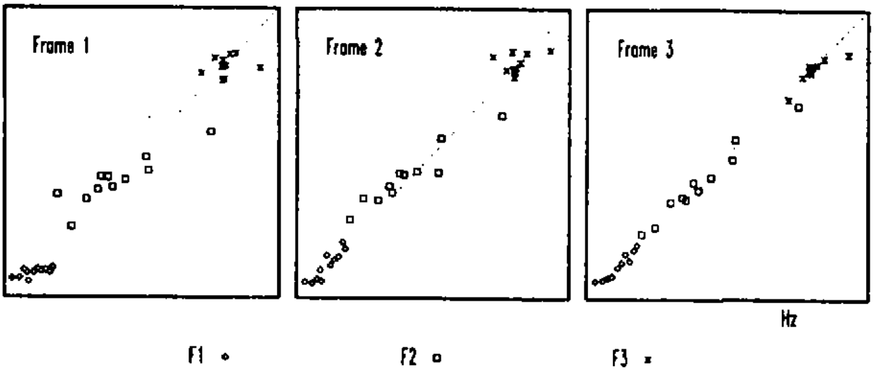
The duration of the transitions between consonants and vowels for the RP speaker studied here was found to vary systematically with the *length of the vowel*, the *voicing* characteristics of stops and fricatives, and with the *manner of articulation* of the consonant. These variations should be taken into account when generating the transitions between adjacent segments.

Coarticulation

It has been known for some time that formant values will be modified for vowels [3,4] and for consonants [5,6] in accordance with the immediate segmental context. We next investigate how formant values change in the /CV-/ combinations from our own RP corpus. In a three-way fixed effects ANOVA (vowel×formant×context) for vowels produced in isolation compared with the average formant values for vowels produced in /CVC/ syllables, context did not produce a significant effect nor enter into any significant interactions. It would appear that on average context does not affect the formant values for this speaker. In view of the results in the studies cited above, this is rather surprising. More natural recording material might have produced significant shifts in vowel formant values. Alternatively, stressing conditions and duration may be the main factor for this speaker in relation to formant variability. For the present, we conclude that it is not necessary to modify vowel formant target values in context.

Consider next whether any relationship obtains between formant values measured at frames <1>, <2> and <3> relative to frame <4>, which represents the vowel steady-state. Figure 2 shows the formant centre frequencies for F1, F2 and F3 in each of the frames <1> to <3> plotted as a function of the corresponding formant frequencies in frame <4>, the vowel steady-state, for a bilabial stop environment[1]. The principal diagonal has been included in each plot for clarity. Each data point represents the formant values for each of the eleven RP monophthongs.

Figure 2: Formant frequencies for the consonants [p,b] and during the transition as a function of the corresponding values for the vowels



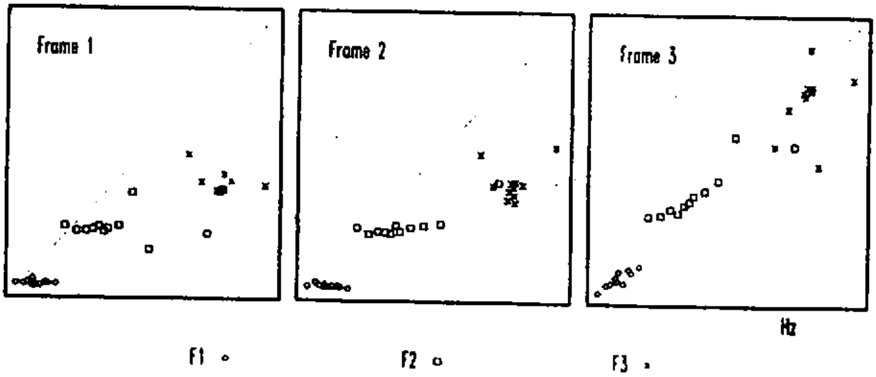
Let us consider these plots in detail. If formant values do not correlate between frames <1>, <2> and <3> relative to frame <4>, then we might expect there to be little influence from the following vowel. Variation in formant frequency is therefore non-systematic. If there is a strong correlation, however, then the influence of the vowel is evident. If formant values coincide with the principal diagonal, then the target values for the vowel have already been attained. In the case of the bilabial stop (Figure 2), a different relationship obtains between each of the three formants in frame <1>, here the release of the consonant, and in frame <4>, the vowel steady-state. F1 for the consonant does not change much regardless of the F1 values for the vowels. Similarly, F3 varies unsystematically relative to the vowel F3. F1 and F3 are therefore insensitive to the nature of the following vowel. By contrast, F2 for the consonant seems to correlate with F2 for the vowel ($r = 0.90$, $p < 0.01$). The range of F2 values for the consonant, however, is rather less than for the vowel and if a least-squares line is fitted to the data, it is significantly different from the principal diagonal ($t[9] = 6.217$ for the intercept and -6.509 for the slope,

---

[1] Data for voiced/voiceless cognates were pooled, since there were no significant differences between the formant frequencies across the voicing distinction.

p < 0.01 in both cases). The second formant values of the consonant are therefore related to those for the vowel even at the onset of the syllable, though they are not identical with the vowel target values. The F2 locus for bilabial stops is therefore highly sensitive to vocalic context. Moving to frame < 2 >, the third formant still varies independently of the vowel. The first formant, however, correlates strongly with the first formant of the vowel steady-state ($r = 0.91$, $p < 0.01$). Indeed, the parameters of the least-squares line fitted to these data do not differ significantly from the principal diagonal ($t[9] = 0.304$ for the intercept and -0.164 for the slope). F2 again shows a strong positive correlation ($r = 0.95$, $p < 0.01$), but the least-squares line still differs from the principal diagonal ($t[9] = 6.094$ for the intercept and -5.817 for the slope, $p < 0.01$). Finally, in frame < 3 > all three formants show strong correlations ($r = 0.95$ for F1, 0.98 for F2, and 0.87 for F3, $p < 0.01$ throughout). The first and third formant values lie along the principal diagonal, suggesting that the target values for the vowel have already been attained at vowel onset. There is still some deviation between the least-squares line fitted to the data for F2 and the principal diagonal ($t[9] = 3.532$ for the intercept and -3.571 for the slope $p < 0.01$). The effect of the consonant is therefore still evident at vowel onset. Where a strong correlation obtains, the parameters of the least-squares line could be used to generate appropriate formant values for the consonant based on the formant values for the vowel. Where no correlation appears, the mean formant value for a given consonant can be used.

Figure 3: Formant frequencies for the consonant [r] and during the transition as a function of the corresponding values for the vowels



Consider now the formant values associated with [r]. Figure 3 shows similar plots to Figure 2 above for the alveolar approximant. The data appear less homogeneous than for the bilabial stop. With only one speaker, however, it is not yet clear how this may be resolved. Nonetheless, tendencies do emerge which may be used for the cosegmentation rules. For frame < 1 >, no correlation obtains for any of the formants. There is apparently no evidence of the following vowel in the formant structure for the consonant at this stage. By frame < 2 >, however, the second formant values for the consonant correlate with those for the vowel ($r=0.87$, $p<0.01$). The slope and intercept of the least-squares line fitted to these data are significantly different from the principal diagonal as might be expected ($t[9]=-11.641$ for the slope and 6.639 for the intercept). The first and third formants still show no correlation (although the correlation for the first formant is significant at the 0.05 level). The final frame reveals strong correlations for the first two formants, but not for F3 ($r=0.89$ for F1, and 0.90 for F2, $p<0.01$ for both). The parameters of least-squares lines fitted to the data for F1 and F2 differ from the principal diagonal (for F1 $t[9]=2.912$ for the intercept, $p\leq 0.05$, and -4.022 for the slope, $p<0.01$; for F2 $t[9]=4.230$ for the intercept and -4.333 for the slope, $p<0.01$ for both). Generating the formant values for [r] therefore involves a different strategy than for bilabial stops. The influence of the following vowel is already present at the syllable onset (the burst release) for the stops [p,b], and by the onset of the vowel, F1 and F3 have already reached their target values. By contrast, the influence of the following vowel on [r] appears only at the offset of the consonant, and the effect of the consonant on the formant values is still evident at vowel onset.

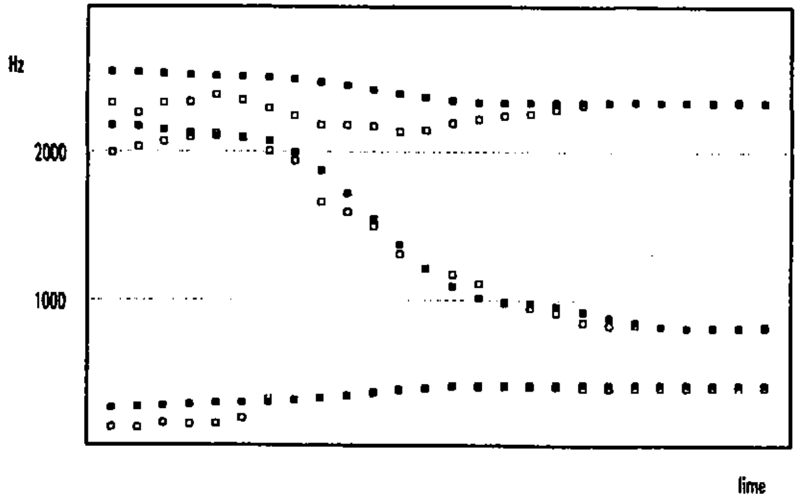
This kind of analysis was applied to the other consonants. We may summarize the general results as follows. At syllable onset, the notional steady-state portion of the consonant (frame < 1 >), no influence is apparent from the following vowel for the consonants [t,d,k,g,f,v,θ,ð,s,z,ʃ,r,m,n,w,j] (the correlations for the second formants of [k,g,f,v,θ,ð,j] were significant at the 0.05 level only), while for [p,b,l] the vowel is already contributing to the value of the second formant. At consonant offset (frame < 2 >), there is still no influence on the formant values from the vowel for the consonants [s,z,n,w]. For [j], F1 alone shows some effect (the correlation for F1 of [r] is significant at the 5 per cent level only), and for [t,d,f,v,θ,ð,ʃ,l,r,m], the following vowel affects F2 alone. For the bilabial stops [p,b], both F1 and F2 are influenced by the vowel; and for the velar stops [k,g], all three formants are affected. By vowel onset, the third formant of [ʃ,l,r] shows no correlation (the correlations are significant for [ʃ,l] for $p\leq 0.05$ only). All other consonants show some vocalic influence to all three formants. For [p,b,m,n,w,j], F1 at vowel onset has reached the vowel target; for [m,n,w], F2 is at the vowel target; and for [p,b,k,g,j], F3 is at the vowel target. For the alveolar consonants [t,d,s,z] and for the labiodentals [f,v] and the dentals [θ,ð], all three formants are still affected by the consonant at vowel onset; for the velar stops [k,g] and the lateral [l], F1 and F2 differ from the vowel target values; for the glide [j], F2 is not the same as the vowel target; and for the other glide [w] and the nasals [m,n] , F3 alone has not attained the vowel target. Inasmuch as the influence of these consonants is still significant at vowel onset therefore it can only be with respect to the third formant.

Some common tendencies have emerged. The bilabial and velar stops seem to show the most vocalic influence during what we might wish to refer to as the consonant in /CV/ combinations. Additionally, [p,b,m,n,w,j] seem to have least influence on vowel formants at vowel onset, while the fricatives [f,v,θ,ð,s,z,ʃ] and the alveolar stops [t,d] maintain their effect upto this point[2]. As more data become available, clear patterns may well emerge. For the present, it is apparent that context specific formant values for consonants may be generated with reference to vowel target values and some linear transformation depending on the consonant in question.

Cosegmentation

How are these findings integrated into the set of cosegmentation rules? In cases where no vowel/consonant dependency emerges, formant parameters based on the mean value across all instances of a given phone are stored in a formant library. Now, suppose we want to generate the /CV/ combination [jɔː]. The formant parameters for the vowel steady state are retrieved from the library. None of the three formants for [j] correlates with the corresponding vowel formant at the notional consonantal steady-state, so the formant values for the consonant are also all obtained from the library. At the offset of the consonant, F2 and F3 were not affected by the vowel and appropriate values are taken from the formant library. F1, however, shows the effect of the neighbouring vowel, and a context specific value is calculated. At vowel onset, F1 and F3 have attained the

Figure 4: Measured and calculated formant values for [jɔː]



---

[2] This echoes some of the findings of Stevens *et al.* [6], namely that alveolar consonants are less prone to influence from the following vowel than other consonants, possibly on account of the articulatory gesture required for the alveolars.

appropriate vowel target values. F2, however, must be calculated. We now have the four frames representing the consonant, the consonant offset, the vowel onset and the vowel steady-state. Next we turn to the temporal organization of the segments. The duration of the consonant and vowel are given (in this case 50 msec for the consonant and 90 msec for the vowel), but the duration of the transition must be calculated. The mean transition duration for glides is 82.7 msec; [ɔː] is a phonologically long vowel, so that this value must be altered by a factor of (61.3/52.5), which gives approximately 96.6 msec (in practice rounded to 100 msec). Formant values between the four basic frames are interpolated using a half-cosine across the required number of frames. Figure 4 shows the original measured formant values for [jɔː] as spoken by our speaker (open symbols) and the values automatically generated values (filled symbols).

Conclusion

The duration of the transition in a range /CV/ combinations and formant centre frequencies at various stages throughout the temporal progression from consonant to vowel have been systematically analysed. The transition duration was shown to be dependent on the *voicing* and *manner of articulation* of the consonant and on the *(phonological) length* of the vowel. The relationship between formant values during the consonant and vowel reveal how far into the adjacent segment the influence of a different segment may extend. The implications of this for synthesis are clear. The results of the analyses detailed above form part of a preliminary set of cosegmentation rules for the IBM text-to-speech system. The implications for perception, however, are only recently beginning to be investigated [7]. It may be possible in the future to integrate the current acoustic work into the perceptual framework Fowler suggests, thereby linking production and perception in a meaningful way.

References

[1] P.R. Alderson, G. Kaye, S.G.C. Lawrence and D.A. Sinclair, 'A Speech Analyser based on the IBM Personal Computer', Proceedings of the Institute of Acoustics, Vol. 6, Part 4, 235-239, (1984)

[2] 'Speech Processing at the UKSC', Internal IBM UKSC Report

[3] B.E.F. Lindblom, 'Spectrographic study of vowel reduction', J. Acoust. Soc. Am. (35) 1773-1781, (1963)

[4] K.N. Stevens and A.S. House, 'Perturbation of Vowel Articulation by Consonantal Context: An Acoustical Study', J. Speech Hearing Res. (6) 111-128, (1963)

[5] S.E.G. Öhman, 'Coarticulation in VCV Utterances: Spectrographic Measurements', J. Acoust. Soc. Am. (38) 151-168, (1965)

[6] K.N. Stevens, A.S. House and A.P. Pauli, 'Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation', J. Acoust. Soc. Am. (40) 123-132, (1966)

[7] C.A. Fowler, 'Segmentation of coarticulated speech in perception', Perception & Psychophysics (4) 359-368, (1984)

Proceedings of The Institute of Acoustics

MORPHOPHONOLOGY IN THE CSTR TEXT-TO-SPEECH SYSTEM

Linda Shockey

Centre for Speech Technology Research, University of Edinburgh

Because the text-to-speech system that we are in the process of building is predicated on the use of a morph dictionary, i.e. a dictionary in which are stored meaningful subunits of words rather than the entire words found in conventional dictionaries, we almost immediately decompose input words into their constituent morphs. While this process is extremely useful for a variety of reasons, it presents us with the problem of pronouncing words correctly at output, after the morphs have been re-conjoined.

Morphophonology is the branch of linguistic analysis which deals with changes in pronunciation which occur when morphs are concatenated. A simple example of this process can be seen in the last word in the previous sentence, the 'n' in the prefix 'con' will assimilate to the following sound, to become /ng/. Another example of a morphophonological change is the variable pronunciation of the past morpheme ed, which is pronounced /d/ after most voiced sounds (fired, filled,) but /t/ after most of the voiceless sounds (locked, missed, laughed).

Typically, a linguistic exposition of morphophonology treats as equivalent those alternations which are signalled in spelling and those which are not. This means that the morphophonological literature is of limited use to us, because we have to deal only with those changes not represented in spelling, such as those found in the past tense morpheme mentioned above. For us, alternations such as divide/division or appear/apparent do not need to be related to a single form; there can be two entries in the dictionary corresponding to the two pronunciations. On the following page are some of the simple consonantal variations which our system must allow for.

CSTR MORPHOPHONOLOGY

1. Palatalisation. This is a complex process which can be divided into several separate rules.

    a. morpheme-final /s/ which is spelled with 'ss' goes to /sh/ before morpheme-initial /y/, and the /y/ is then dropped, as in 'pressure' and 'fission'.

    b. morpheme-final /s/ which is spelled with 's' goes to /zh/ before morpheme-initial /y/, and the /y/ is then dropped, as in 'leisure' and 'erosion'.

    c. morpheme-final /t/ preceded by a vowel goes to /sh/ before the suffixes 'ia', 'ion' and 'ial'and their respective initial /y/ is dropped, as in 'inertia', 'motion', and 'spatial'.

    d. morpheme-final /t/ preceded by a consonant goes to /ch/ in the same conditions as (c) above.

    e. morpheme-final /t/ goes to /ch/ before the suffixes 'ure, ute, ury' and 'ual', as in 'feature', 'statute', 'century', and 'mutual'.

2. Pluralisation. This, again, is a process requiring several rules:

    a. the plural morpheme ('s') is to be pronounced /i z/ after the strong fricatives /s,z,sh/ and /zh/, as in 'horses', 'roses', 'wishes', and 'beiges'.

    b. after all other voiced consonants and vowels, it is to be pronounced /z/, e.g. bags, buns, lives, bees.

    c. otherwise, pronounce it /s/, as in 'cats' /k a t s/.

3. Morpheme-final 'r.'

    a. when a morpheme-final 'r' is followed by a consonant, do not pronounce it.

    b. otherwise, do.

4. Morpheme-final 'gn.'

    a. when a morpheme-final 'gn' sequence is followed by a derivational suffix, pronounce it /g n/, as is 'signature, recognition, dignity.'

    b. otherwise, pronounce it /n/, as in 'align, signing, maligned.'

5. Morpheme-final 'mn.'

    a. when a morpheme-final 'mn' sequence is followed by a derivational suffix, pronounce it /m n/, as in 'damnable, hymnal, solemnity.'

    b. otherwise, pronounce it /m/, as in 'hymn.'

This type of alternation is obviously very easy to deal with: one simply looks for a sequence of symbols across a morpheme boundary and alters it to its correct shape.

There is a more complicated set of morphophonemic alternations involving vocalic changes when suffixes are added. An example of this is the pair "tone/tonic." Without morphophonemic adjustment, both members of this pair would be pronounced with the vowel of "tone."

It is well known that there is more than one kind of alternation in vowel quality in English. Elsewhere in our system we deal with changes in vowel quality which occur when stress moves about in a word, causing stressed syllables to take on full vowel quality and unstressed syllables to reduce. The morphophonology component of our text-to-speech system, by

contrast, handles those cases where conjoining morphs causes vowel changes either 1) with no change in stress or 2) where stress placement changes, but the stressed vowel does not assume the expected quality.
An example of the latter is "agile/agility", where the movement of the stress onto the second syllable occurs simultaneously with a vowel change in the second syllable usually associated with unstressed forms.

Following are some examples of the type of alternation we are considering at this point:

Vowel Changes Associated with the Addition of a Derivational Suffix

ity

asinine/asininity
saline/salinity

agile/agility
___ile/___ility

verbose/verbosity
benign/benignity

itude

grate/gratitude
sole/solitude

ic,ical

academe/academic
aesthete/aesthetic
athlete/athletic
metre/metric
type/typical
(en)cycle/encyclical

al,alise,ar

doctrine/doctrinal
urine/urinal
line/linear

ify

mode/modify
type/typify
vile/vilify

Vowel Changes Associated with an Inflectional Suffix

-en forms	-ed forms
shriven	dreamed
thriven	leaned
(a)risen	leaped
chidden	spelled
bestridden	spilled
etc.	

When this research began, it was thought that broad patterns would emerge in alterations of stem vowels with the addition of particular suffixes, so that features could be constructed to be included in the dictionary entries of the stems. These features, it was reasoned, would allow us to use only a small set of generalised transformations to cover all morphophonological variations. For example, we assumed that there would be a class of stems containing [ai] which alternated with [i] when a particular suffix or class of suffixes was added, as in the pair "line/linear." If such a class existed, a feature such as "class2" could be included in the list of features in the dictionary and a substitution could be effected whenever the triggering combinations for Class 2 were found. This would allow us to put "line/linear" in the same alternation class as "child/children" and "thrive/thriven." In fact, two factors were found to stand in the way of such an approach. First, the number of stems which changes with the addition of a particular suffix is, on the whole, quite small. The stem "line" probably forms a class of one in the way it combines with the suffix "ar." Other stems with the same shape (such as "fine" and "pine") do not combine with "ar," and an invented stem (such as "grine") might or might not be pronounced like "linear" if combined with it. Second, while some morphophonemic changes do occur in stems when suffixes are added as in "cave/cavity," the great majority of effects are across suffixes, as in "divine/divinity" and "agile/agility." To further complicate matters, what seems to be the same rule for changing vowels can operate both between stem and suffix and between two suffixes, for example:

Stem + Suffix		Suffix + Suffix	
sane + ity =	sanity	hum + ane + ity =	humanity
[ei]	[a]	[ei]	[a]

In the case of the stem + suffix, however, we are dealing with a few sporadic cases, whereas the suffix + suffix change is perfectly regular.

Based on this, we think that there are some cases where our originally-conceived solution is feasible, though the features must be more specific than we had imagined. All of the stems which change with given suffixes can be listed in the dictionary with features such as "ei_a_iti," specifying that the stem vowel [ei] changes to [a] when it is found with the suffix "ity." These features, then, are really a set of tests which specify whether a change occurs.

Since the suffix combinations have predictable pronunciations, they can be handled by rule, much like the simple consonantal combinations cited above. The sequence 'suff(ane) + suff(ity)' will always be rewritten as [aniti], in other words.

This solution does not allow us to make the generalisation that some changes are the same whether they are between stem and suffix or suffix and suffix. But it may be that such a generalisation would be false, and that we are in fact dealing with two different processes, one nonproductive and one productive, which yield the same result.

In summary, we see morphophonemic processes as falling into the following general classes:

A. Consonantal changes which are easily handled by rule.

B. Vowel changes:

1. in stems, which are triggered by one suffix or a combination of suffixes. These we handle in the dictionary, because they are sporadic.
2. in suffixes, which are triggered by following suffix(es). These we handle by rule, because they are predictable.