Synthesising Spontaneous Speech

W. N. Campbell

ABSTRACT
This chapter addresses the issue of producing synthetic speech for an interpreted dialogue where the emotional content of the original utterance is to be preserved; it describes differences in speaking style between read and spontaneous speech from the viewpoint of synthesis research and discusses the development of a synthesis system incorporating labels to encode the prosodic and segmental variation. Spontaneous speech confronts us with phenomena that were not encountered in corpora of prepared or read speech, and to account for these we are increasingly having to identify higher-level units of discourse structure and speaker involvement.
The chapter makes three specific claims: (a) that it is not necessary to label or predict fine phonetic detail in order to be able to produce natural-sounding speech, *i.e.*, that the distinctive characteristics of phonetic detail are determined top-down from the prosodic environment, which is itself dependent on the speaking style. (b) that the labelling of segmental and prosodic characteristics such as are required for the synthesis of non-interactive speech can be done adequately for speech synthesis using automatic techniques, leaving the human labeller free to identify higher-level discourse-related aspects of the speech. (c) that instead of minimising the size of the source database of speech units, we should instead be concerned to maximise its variety and to efficiently select from it the units that most closely express the characteristics of the target speech. The ChATR resynthesis toolkit performs many of these tasks.

1 Introduction

Speech synthesis is not spontaneous, nor can it be. However, there are applications of synthesis where modelling of the spontaneous characteristics of natural speech is required, such as in an interpreted dialogue where speakers talk in their own language and the speech is then automatically converted into the language of the listener. In such a dialogue the prosodic attributes, such as speed of speaking, degree of segmental reduction, tone-of-voice, etc., carry information that signals amongst other things the speaker's mood, commitment to the utterance, speech-act type, and stage of the discourse. For the successful interpretation of such information, the system must be

A key point in the work being presented here is to show that units for synthesis can instead be excised from a corpus of speech produced in less constrained situations, which therefore includes more natural prosodic variation typical of different styles of speech. Because of the variation in such a corpus though, the accuracy of the labelling becomes much more important as it becomes necessary to identify and select source units not just from an appropriate phonemic environment but also with respect to the prosodic and voice-quality dimensions as well. If by dint of improved labelling we can extract the units for concatenation from a context that is similar to the target in all significant dimensions, then we can reduce the amount of signal processing that will be required to produce the appropriate intonation, and therefore maintain a level of synthesis that is closer to the quality of real human speech. In this way we shift the main task of synthesis research away from the *modelling* of speech and in the direction of its characterisation (or *labelling*) instead.

Most of the corpora so far studied for speech synthesis have been of read speech. There is already a considerably body of experience in the automatic or assisted labelling of segmental and prosodic aspects of such corpora ([Tal94, Cam92, Cam93, Cam92, Wig95, Kie95, Koh94, Bru96]), but for the synthesis of dialogue or conversational speech, then such additional aspects as voice quality, hesitations, and speaking style will also need to be identified as additional features. However, rather than resulting in a proliferation of the number of labels that are required, this actually reduces the labelling load. We will see below that the labelling can be performed in a simple hierarchical way, with each level inheriting features from higher-level descriptors, so that rather than describing (and having to identify) minute variations in articulatory characteristics we can predict their occurrence instead. That is, by labelling the higher-level features of a spoken utterance we are thereby able to predict the circumstances under which the dependent lower-level characteristics change.

1.2 Natural speech

Speech is 'natural', but not all speech is similarly natural, and recorded speech, especially when recorded in a controlled environment as a source for synthesis units, can be highly constrained. In its natural form, speech is inter-personal and often functionally goal-directed, but in recordings of 'lab speech', where the listener is replaced by a microphone, the speech becomes production-based rather than listener-oriented, and there are significant and perceptually relevant differences between a spontaneous natural utterance and a prepared one that mimics it even though the text of what is said may be identical [Bla95]. As a consequence, the materials that in the past we have collected. analysed, and ultimately synthesised from may not be representative of what people actually *do* when they speak normally.

Furthermore, because the reader producing a source-unit database for

synthesis is faced with the daunting task of having carefully to read into a microphone long lists of unconnected sentences (or worse, nonsense-words) to produce all the required sound combinations, there is a high probability of boredom or fatigue having effects on the voice quality in such recordings.

To obtain source units for the simulation of lively natural speech, it may be preferable to replace production controls at the recording stage with statistical controls in a later analytical stage, and to use these to process instead large representative corpora of spontaneously produced spoken material. Such corpora are now becoming more widely available but the tools for their analysis were developed for a more restricted speaking style, when read speech was the main source of data.

2 Labelling speech

Traditional phonetics labels speech according to segmental content alone, and while allowing the use of diacritics to describe prosodic variation, typically regards this as very much a secondary feature. I argue the contrary: that in order to describe a speech segment sufficiently to use it in concatenative synthesis where the goal is to reproduce the characteristics of natural speech, we have to label *both* the segmental and the prosodic attributes equally. This is because simple phonetic labels do not sufficiently describe the location on the scale of hypo- hyper- articulation of segment sequences with supposedly identical phonemic structure.

A serious consequence of this under-labelling is that units with the same phonetic labels taken for synthesis from naturally occurring speech may not be similar enough to concatenate without a noticeable discontinuity. This explains why such care has been taken to reduce the variation in typical source databases for synthesis. Worse, in the proposed synthesis of interactive speech, the message being given by the manner of articulation may not conform to the intended interpretation of the utterance, and we may end up synthesising with the 'wrong tone-of-voice'.

Lindblom [Lin90] has shown that the phonatory characteristics of articulation vary according to speaking style and speaker familiarity. Kohler has similarly described a cognitively-based reduction coefficient [Koh95, Koh96] under the control of the speaker that governs reduction and elision, causing scalar variation in the articulation of a given sequence of phones in different contexts. Mechanisms for these effects have been described in articulatory phonology in terms of overlapping constituents [Col92] but also as intentional [Wha90]. Since they are predictable from higher-level features of the discourse, such as speaking rate, then it should be sufficient to know the speaking style (and its prosodic correlates) in order to describe the degree of reduction on any given segment in an utterance.

For identifying such higher-level features, canonical segment labels can first be automatically aligned in order to provide access to discrete portions

of the speech waveform from which we can then extract prosodic details. These in turn can be used to detect and label the higher-level structural and stylistic features which will later be used to account for the finer articulatory differences that are then predictable from context,

2.1 Automating segmental labelling

What can be predicted does not need to be explicitly labelled. Kohler (this volume) argues that a linear segmental representation of canonical citation forms can account well for the phonological reorganisation of speech, and shows that although a segment may be elided or deleted in the production of fluent speech, a non-segmental residue remains to colour the articulation of the remaining segments. This supports our contention that rather than attempt a fine labelling of the surface representation of sounds in an utterance, it is preferable to label only the underlying canonical segment sequences but to relate them to their prosodic environment separately in a multi-tiered description. Similarly in synthesis, rather than predict the microsegmental variation, we can use selection according to prosodic environment to bypass this difficult task. A canonical representation of the phone sequence is easily accessible from a machine-readable pronunciation dictionary, so given an orthographic transcription of a speech corpus, segmental labelling can be automated to a large extent by using speech recognition technology to predict and align a default phone sequence. This is then complemented by an encoding of the prosodic structure of each utterance to capture the interactions.

By training single-phone hidden Markov models (HMMs) corresponding to the set of phonetic labels in a machine-readable pronunciation dictionary, and generating networks of default pronunciations for each word in the orthographic transcription, we can obtain a first-pass estimate of the segmental realisation of each utterance. Separate lexical sub-entries must be included for some particularly different pronunciation variants such as 'gonna' for 'going to', but in general a single pronunciation for each word will suffice[1]. A finer segmental alignment can then be achieved after a second pass by using Baum-Welsh re-estimation [HTK93] to retrain the HMM models specifically for each corpus, using the transcription derived from the orthography to constrain the alignments. We can thereby achieve segmentation accuracy comparable to a human transcription (see for example [Tal94]).

A criticism of this blind transcription technique is that without human intervention we do not know for certain what pronunciation a word was given in a particular utterance, and that it is possible for example that the

[1] In the synthesis stage, only one pronunciation for any word will be generated, but its actual realisation will depend on its prosodic context.

consonant cluster in 'handbag' although actually pronounced as [mb] was force-aligned (from the canonical form) as an [ndb] sequence. The counter to this criticism is that given the supplementary information about the prosodic environment, such knowledge is no longer required; the same sequence in a markedly prominent or contrastively focussed environment is likely to be given one pronunciation, and in a normal or reduced environment the other, or something in-between. The claim being made in this paper is that the degree of segmental reduction is predictable from the prominence marked on the sequence in conjunction with the speaking style.

2.2 Automating prosodic labelling

Whereas prosodic variation is scalar and multi-dimensional, including at least fundamental-frequency, segmental duration, and amplitude changes, prosodic *structure* can be represented as binary and in two dimensions by a combination of the higher-level labels of prominence and phrase-finality, as in the ToBI system of prosodic transcription [Sil92]. In read speech at least, phrasal boundaries and prominences appear to be the most basic elements marking prosodic structure, and we can predict much about the phonatory (acoustic) characteristics of a segment from knowledge of its place in the syllable and of that syllable's position with respect to its neighbours the various levels of prosodic phrasing and prominence.

Taking segmental duration as an example, figure 1 (see [Cam93]) illustrates three types of prosodic context that affect the duration of a syllable. In terms of lengthening, the effects of prominence are biased more towards early segments (onset and peak) and those of phrase-finality on later ones (offset or ryme). Rate-related lengthening affects segments more uniformly. A syllable immediately before a prosodic phrase boundary is likely to be lengthened, with amplitude low and decaying, and it may exhibit vocal fry in the ryme. The lengthening is likely to be greater with increasing strength of phrase break, and a pause is likely to follow if utterance-final. There will also be lengthening observed in a prominent (or nuclear accented) syllable, but in this case it is likely to be more marked on the onset segments [Jon95, Cam93] and there may be more aspiration after plosives in the onset, increases in spectral tilt resulting from changes in vocal effort [Pie92, Cam95, Gau89, Slu93] and differences in supraglottal phonation arising from local hyperarticulation [Lin90].

In labelling the source database, each syllable is therefore tagged according to the following features to determine its prosodic environment:
(a) $\pm prominent$ (a binary indicator)
(b) $\pm phrase-final$ (binary at three levels of phrasing)
where, 'prominence' on a syllable is defined perceptually as 'having been uttered with a greater degree of vocal effort than surrounding syllables' (and as such frequently but not necessarily co-occurs with lexical stress), and

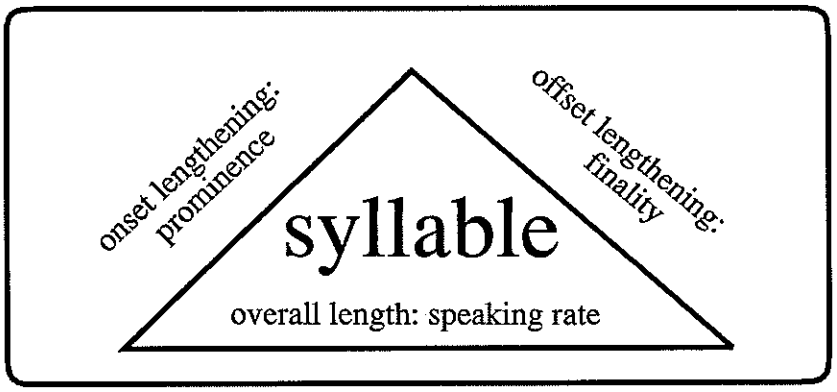

FIGURE 1. The prosodic lengthening effects on a syllable

'prosodic phrase-finality' is defined at three levels: (i) the accentual (minor) phrase, (ii) the intonational (major) phrase, and (iii) its utterance-final variant. Higher levels of chunking will be required (*e.g.*, at the paragraph level for read text and to include disfluencies or turns in more natural speech) but cannot yet be performed automatically.

Wightman & Campbell [Wig95] were able to correctly predict most of the hand-labelled prominences and intonation boundaries in a corpus of forty-five minutes of professionally-read American radio-news speech. Section f2b of the Boston University Radio News Corpus [Ost05] was produced by one adult female speaker and exhibits a consistent marked style typical of professional announcer speech. The corpus had been prosodically labelled by hand according to the ToBI conventions to differentiate high and low tones at intonational boundaries and on prominent syllables, and to mark the degree of prosodic discontinuity at junctions by break indices between each pair of words.

A set of acoustic, lexical, and segmental features derivable from the phone labels, the dictionary, and the speech waveform, was defined and achieved automatic detection of 86% of hand-labelled prominences, 83% of intonation boundaries, and 88% correct estimation of break indices (at ± 1). The acoustic features extracted from the speech waveform for the autolabelling of prosody include (in order of predictive strength) silence duration, duration of the syllable ryme, the maximum pitch target[2], the mean pitch of the word, intensity at the fundamental, and spectral tilt (calculated from the harmonic ratio). Non-acoustic features included end-of-word status, polysyllabicity, lexical stress potential, position of the syllable in the word, and word-class (function or content only). These latter were all derivable

[2]Pitch targets were calculated using Daniel Hirst's quadratic spline smoothing to estimate the underlying contour from the actual f0.[Hir80]

capable of recognising and expressing subtle prosodic and voice-quality changes. It is probable that when using such a system, speakers will be more careful than usual to control their style of speaking, and it is perhaps questionable whether the synthesised translation should be required to sound completely natural (if that were possible) because of accountability issues, but we are instead concerned here with the still theoretical issues of how to automatically identify and label such stylistic information and with the techniques of synthesis best used to encode it.

1.1 Synthesising speech

There are three primary methods of synthesising speech; (a) articulatory synthesis, which produces a speech waveform by modelling the physiological characteristics and excitation of the human vocal tract, (b) formant synthesis, which directly models the acoustics of the speech waveform, and (c) concatenative synthesis, which uses pre-recorded segments of real speech to construct a novel utterance. For the manipulation of prosody, (b) offers the most flexibility, and (a) the most natural built-in constraints, but (c), while producing the most natural-sounding speech, is the most difficult. Because concatenative synthesis employs digitised segments of recorded speech, it reproduces the fine variation of detail that is still too complex or too subtle for the other methods to model, which is why it sounds so similar to human speech, but in manipulating the prosody of a concatenated sequence of segments, we must resort to signal processing and encoding techniques such as [Mou93] that inevitably introduce some distortion and reduce the naturalness. The more the prosody is varied from that of the original recording, the more the waveform is distorted from its natural shape, and the more artifacts are introduced by the processing. By increasing the size and variety of the segment inventory, this problem with concatenative synthesis can be greatly reduced.

Although concatenative synthesis is the most natural-sounding of the three methods, we have yet to hear automatically generated synthetic speech that can consistently be confused with a human original; *i.e.*, speech synthesis has yet to pass its Turing Test. A likely reason for this is that although the source segments (*units*) for concatenation were originally produced by a human speaker, they have typically been excised from recordings of carefully prepared read speech, and although they may be *phonemically representative* of the sound combinations of a given language, they are *prosodically constrained* and invariant, *i.e.*, they form a set of typical sound sequences that represent the phonetic-context-dependent allophones required to reproduce a spoken language, but they fail to adequately model the range of prosodic-context-dependent variation that occurs when speech is produced in various natural contexts. In warping them to a different prosodic configuration, as is required in the synthesis, the original naturalness is lost.

directly from the dictionary used in the aligning.

3 Synthesis in ChATR

ChATR[3] [Cam92, Cam94] is a set of tools that take an arbitrary speech corpus, with its orthographic transcription, and automatically generates from this a labelled database of segments with derived features. Selection from this database is then in terms of a weighted combination of the segmental and prosodic features to satisfy a target utterance specification.

Using simple waveform concatenation, the method is speaker-independent. It is also language independent since the target description for any novel utterance must be completely specifiable in terms of the segmental and prosodic labels of the database from which it is to be generated. The language-specific processingrequired to predict the appropriate representation of the phone sequence and prosody for a text-to-speech synthesiser is not addressed in this chapter, as we take such a representation as basic input to the synthesis module.

From databases prepared in the above manner, we can now produce synthetic speech to reproduce the voice and speaking style from any available speech corpus. Preparing a new speaker (*e.g.*, from 40-minutes of phonetically-balanced utterances) can be completed in less than a day, from initial recording, through segmentation and weight training, to eventual synthesis[4].

The method is *not* speaking-style independent, and we can only model the style(s) of speech found in the source corpus - *i.e.*, news speech always sounds like news speech - but this can be an advantage: with enough disk space, we can now reproduce the characteristics of any speaker or speaking style, given a sufficient source corpus.

Many previous methods of speech synthesis were limited by machine and memory size, and so were constrained to modelling intelligibility rather than naturalness. However, with the advent of multi-media computing, many more resources have become available. The recently-agreed magnetic-optical standard of 4.7 gigabytes for a 'floppy' disk allows sufficient room for more adventurous techniques of speech production, since even a high quality recording (without compression) requires only about a megabyte of memory per minute of speech, and for non-interactive speaking (read-speech), 20 minutes currently seems to be an adequate minimum size.

Once the prosodic and segmental features are labelled for a given database, training of the weights to determine the strength of contribution of any

[3] Collective hacks from ATR (pronounced 'chatter' for obvious reasons).

[4] We have currently tested this process with corpora from twelve speakers of Japanese, five of English, two of German, and (without requiring any changes to the c-code) one of Korean.

given feature in a specific database is performed automatically by jack-knife substitution, removing each utterance of the original database in turn and synthesising an approximation of it using the segments remaining in the database according to a range of different weight settings to produce a measure of the Euclidean cepstral distance for each [Bla95].

Campbell & Black [Cam94] reported results using the BU Radio News corpus as the basis for a resynthesis test of the assumption that labels of prosodic and canonical segmental context suffice to encode the lower-level spectral and articulation characteristics, employing the ChATR speech synthesis toolkit to select segments from a labelled corpus for concatenative synthesis as described above. Using similar jack-knife substitution, we resynthesised each utterance by concatenation of segment sequences selected from the remaining utterances according to suitability of their prosodic environment, with no signal processing performed on the concatenated sequences. Measures of Euclidean cepstral distance between target and synthesised utterances confirmed that the use of prosodic features in the selection resulted in a closer match between the spectra of target and synthesised utterances. When equivalent tokens from the same segmental sequences were selected from less appropriate prosodic environments, ignoring the weights on the non-segmental features, the resulting synthetic speech showed considerable degradation. Table 1 shows similar results for a database of Japanese speech.

Because the source corpus typically includes natural non-speech noises, these can also appear in the synthesis if in an appropriate context for selection. It frequently happens that a sequence of segments across a prosodic phrase boundary is resynthesised using tokens selected from pre- and post-pausal locations such that the 'silence' between them includes an appropriate sharp intake of breath. Such noises coming from a synthesiser make the resulting speech sound even more 'natural'.

4 Spontaneous speech

The range of prosodic variation is much greater in spontaneous speech. As an illustration of the contrasts between read and spontaneous speech in British English, we can examine the durational characteristics shown in figures 2 – 5, which plot mean segmental duration against the coefficient of variance (*i.e.*, the standard deviation of the durations expressed relative to the mean) for each phone class for each speaking style.

The data examined in this section come from four related corpora. The first contains citation-form readings of 5000 English words; the second, a subset of these words in the form of 200 meaningful sentences read as isolated words; the third, the same sentences read in connected form as meaningful sentences; and the fourth, 20 minutes of spontaneous interactive monologue (*i.e.*, dialogue with a passive partner). They are of British

English from a young adult female speaker, and show a wide range of production variation.

We can see from Figures 2 and 3 that in the isolated-word citation-form readings, there is a good dispersion in the mean durations for each phone class, and relatively constant variance in their durations. Figure 4 shows the opposite to be the case for the exact same sequence of words as for Fig. 3 but read in continuous sentences. Here the variance increases and there is considerable shortening so that segments are no longer as distinct in their mean durations. Separate examination of segments in word-initial and word-medial position confirmed that this is not just a result of more phrase-final lengthening (isolated words being also complete phrases) rather, the articulation of the citation-form words was generally slower and more distinct.

When the speech contains little contextual information, and the speaker is concerned to be clearly understood, then segmental durations appear to be maximally separated, exaggerating the differences between the phone types, but as the style becomes more natural and the listener can rely on prosodic phrasing to aid in the interpretation of the speech, then we find more variance in the durations and less distinction between their means; all words tend to be shorter overall and more varied than in the citation-form readings.

The spontaneous monologue from the same speaker, in Figure 5, shows the same trends more clearly. we find not only that the mean durations for all segment types are low and uniform, but that the variances are huge - notice especially that the vertical scale of Figure 5 is *twice* that of the other three figures. When a listener is present, the speaker has a better understanding of the extent of their mutual understanding, and can hurry through some parts of the speech, and linger before others. There is much greater range of variance in the natural speech than was found in the more carefully prepared 'lab' speech.

If we try to predict the prosodic characteristics of such utterances using training data based on read speech we will be able to predict only a small fraction of the variance observed, since many of the factors now coming into play will not have been considered. If, on the other hand, we are able to predict different local speaking styles, marking areas of confidence or high mutual understanding, then we stand much better chance of determining the durational (and other articulatory) characteristics of the segments they include.

4.1 Spectral correlates of prosodic variation

To confirm that this difference of style is not unique to durations, nor specific to the prosody of one speaker, we can compare changes in spectral tilt associated with prominence and focus in read *vs.* interactive speech. That there are significant correlations between changes in a prosodic dimensions

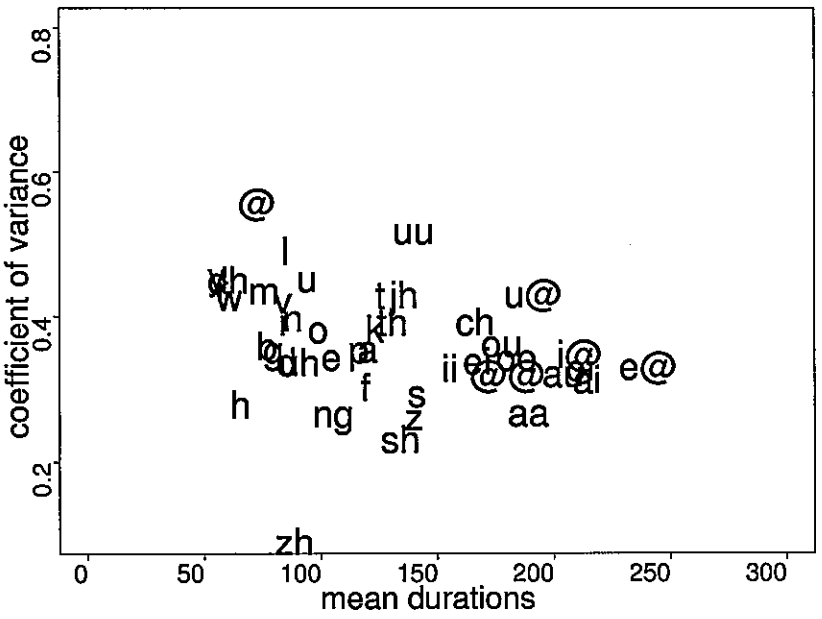

FIGURE 2. Segment durations in isolated-words

and acoustic features of the segmental articulation confirms the validity of this multi-tiered labelling system in describing the variation that occurs in natural speech.

The data used in this section is taken from a corpus of 300 focus-shifting sentences, produced by a young female American speaker, which illustrate the effects of contrastive focus. The sentences were produced in three utterance styles: (a) read in grouped order by set, (b) read in randomised order, and (c) produced spontaneously by elicitation in interactive discourse. Each set of sentences contained syntactically and semantically identical word-sequences that differed only in the emphasis given to each word in different renditions. Shifts of emphasis in the read speech were controlled by use of capitalisation to signal different interpretations, and elicited in the interactive discourse, by (deliberate) misinterpretations on the listener's part.

A study of prominence detection, described more fully in [Cam92, Cam95], showed that speakers change their phonation according to the discourse context and the type of information they impart. The detection algorithm used both duration and spectral tilt[5]. The corpus varied prominence in two

[5]as measured by the relative amount of energy in the top third of an ERB-

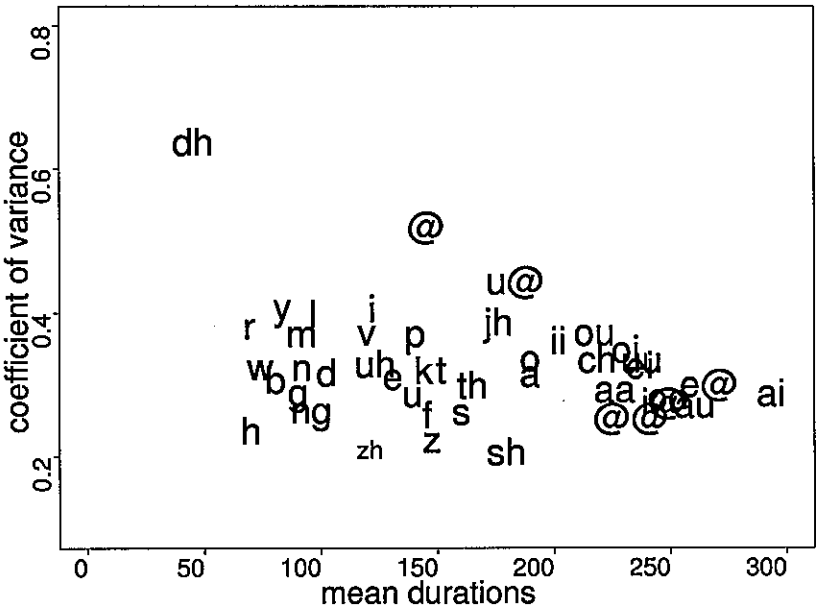

FIGURE 3. Segment durations in isolated-word sentences

ways: (1) by asking the speaker to "emphasise" capitalised words in reading, and (2) by eliciting emphatic corrections of feigned misinterpretations.

Using normalised duration and energy to detect prominences achieved 92% recognition in the clearly read speech, but only 72% in the interactive dialogue (Table 1.). The elicited corrections resulted in a clearer articulation, but durational organisation was more varied, and prominence was not easy to detect automatically using this alone as a cue.

It is interesting to note that although the durational cues to prominence were weakened by greater variance in the interactive speech, the spectral measure was apparently strengthened, as Table 2 shows. We can suppose (like Lindblom [Lin90]) that this trade-off is not coincidental, and that the speaker varies her production according to the needs of the discourse context.

---

scaled spectrum (2kHz – 8kHz).

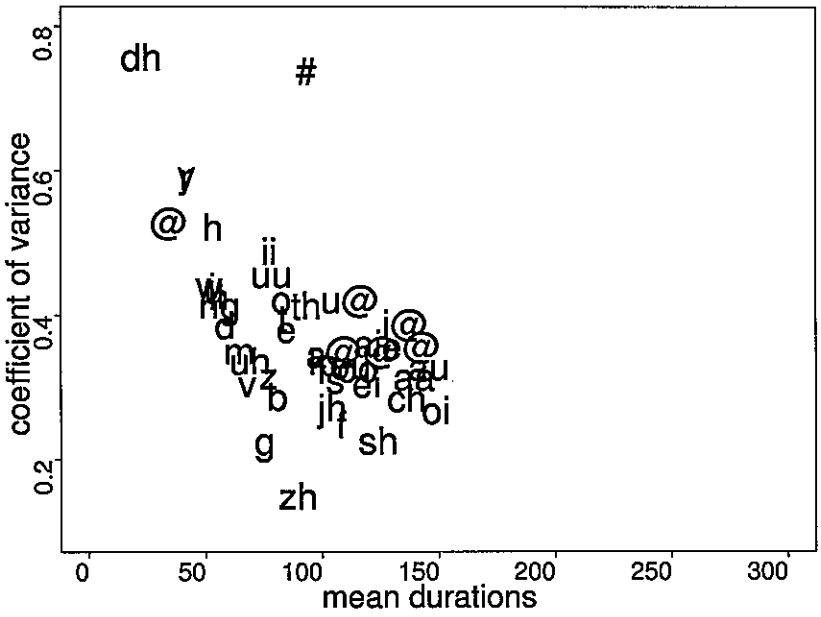

FIGURE 4. Segment durations in continuous sentences

4.2 *Interactive speech*

Although much of the labelling of significant levels of information can now be performed to a large extent automatically and requires only minimal hand-correction for corpora of read speech, these techniques do not extend easily to the processing of dialogue speech or spontaneous monologue. Here we realise the need for extra levels of information to describe the structuring of discourse events that cannot yet be achieved automatically. Whereas the read speech was highly predictable, the unplanned (spontaneous) speech is characterised by bursts of faster and slower sections where the speaker displays switches in speaking style [Bar95], and by much greater variation in f0 range and pausing as she expresses different degrees of confidence, hesitation, involvement, and uncertainty.

In order to compare the speech of one individual, in a highly restricted domain, under a variety of interaction styles, we recorded a native speaker of American English taking one side in a series of twenty task-related instruction-giving dialogues. These were performed in a multi-modal environment, alternatively with and without a view of the interlocutor's face [Fai94].

Transcribing the orthography of such spontaneous speech required more than just the skills of an audio-typist, and to allow auto-segmentation,

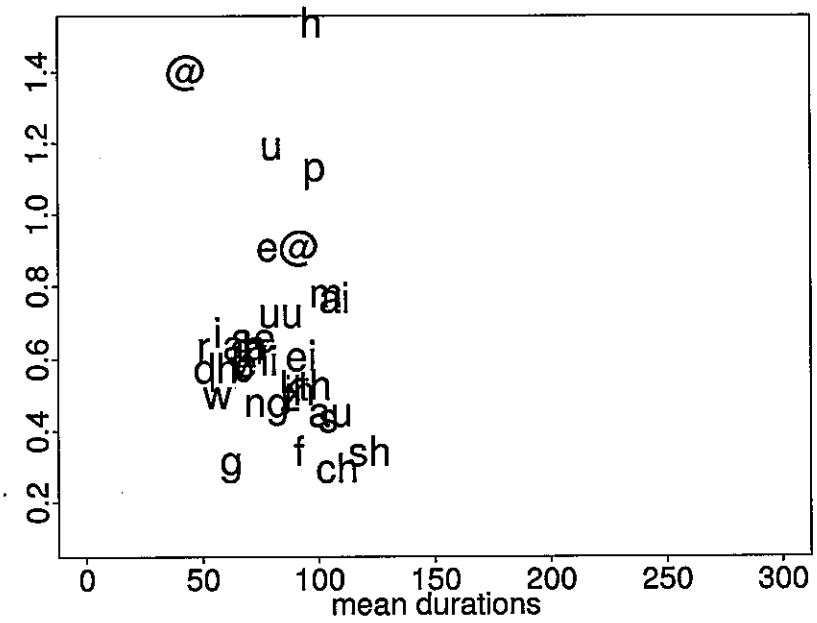

FIGURE 5. Segment durations in spontaneous speech

decisions had to be made about marking disfluencies and repairs. To include this information in our labelling, two extra tiers of information were added to the basic ToBI transcription. One, after Nakatani and Shriberg [Nak93, Shr95], which extended the miscellaneous tier of the ToBI transcription to describe interruptions in the speech flow, and one after Stenström [Str94]), to label IFT (illocutionary force type) speech-act information. The following set was used:

> *inform, expressive, good_wishes_response, apology_response, invite, vocative, suggest, instruct, promise, good_wishes, yn_question, do_you_understand_question, wh_question, yes, no, permission_request, acknowledge, thank, thanks_response, alert, offer, offer_follow_up, action_request, laugh, greet, farewell, apology, temporize, hesitation, confirmation.*

Hirschberg has noted that the major differences between lab speech and spontaneous speech appear to be prosodic (concerning speaking rate and choice of intonation contour) [Hir92, Hir95], but acknowledges also significant segmental differences. She notes for example that some disfluencies in spontaneous speech are marked by characteristic phonetic effects, such as *interruption glottalisation*, which is acoustically distinct from articulatorily similar laryngealisation. In labelling to include these characteristics, we

need not just the prosodic and segmental information derived from HMM alignment, but also an indication of fluency or commitment. That is, we need to label not just what the speaker says but what she is doing in saying it, and how she feels about what she is saying.

As an example, in the dialogue corpus the word 'okay' was said 140 times. It was variously labelled as 'acknowledge', 'confirmation', 'offer_follow_up', 'accept', and 'do_you_understand_qn.', etc., twelve categories in all. The intonation, duration, and articulation varied considerably; sometimes short, sharp, and rising, on a high tone, sometimes slow and drawn out on a falling tone. Since we were able to find significant correlations between the intonation and the speech act label for most of these cases (see [Bla95] for details), we continue in our assumption that instead of trying to predict and model the lower-level acoustic variations, we should instead be accessing them through higher-level labels.

Spontaneous speech appears to be most marked in terms of its rhythmic structuring, exhibiting greater ranges of variation with corresponding differences in phonation style. These prosodic changes appear to have clear correlates in the speech-act labels that we are now using. However, to more fully describe them, we need also to formalise a measure of the speaker's commitment to her utterance. Impressionistic comments such as 'she's thinking ahead', 'her mind's not on what she's saying', 'she's said this many times before', and 'she doesn't quite know how to put this' are triggered by such differences in speaking style, but none of the labels we have considered so far are sufficient to mark such differences. The next step in this work is to determine the appropriate labels, in order to categorise their prosodic and articulatory correlates. Since human listeners can respond consistently to such subtle speaking-style changes, then the clues must be present somewhere in the speech signal but rather than search at the acoustic level, we will continue to explore higher-levels of labelling in an atttempt to capture them.

5 Summary

To summarise the main points of this chapter, I have argued that concatenative synthesis currently offers the best method of generating synthetic speech by rule, and that ordinary speech databases are a better source of speech units for synthesis than are specially recorded databases. The success of selecting units from such a database crucially depends on the labelling of the database.

For the efficient characterisation of speech sounds, it is not necessary to label the fine phonetic features explicitly nor to attempt a numerical description of their prosodic attributes, because these are necessary consequences of the higher-level structuring of the discourse in which they occur. By labelling a large corpus of natural speech as a source of units

for concatenative synthesis and selecting non-uniform-sized segments by a weighted combination of segmental and prosodic characteristics we can reduce the disruptive warping that is required to fit a waveform segment into a predicted context and therefore maintain a higher level of naturalness in the resultant speech.

For non-interactive or read speech, knowing the tri-phone context of a segment, its position in the syllable, and whether that syllable is prominent, prosodic-phrase-final, or both, allows us to predict enough about its lengthening characteristics, its energy profile, its manner of phonation, and whether it will elide, assimilate, or remain robust. In the case of interactive speech, however, a significant part of the message lies in the *interpretation* of *how* it was said, and to encode sufficient information about such aspects of the utterance as voice-quality and speaking style, we need to label discourse and communication strategies that allow the listener to estimate the state of mind of the speaker and her commitment to the utterance.

Furthermore, when a large and sufficiently representative corpus is labelled in terms of the factors that govern phonemic, phrasal, prosodic, speech-act etc., variation, then it will no longer be necessary to attempt to predict the fine details of articulation or prosody at all; it will sufficient to select a segment from a context with the appropriate labels in order to characterise the desired target speech. The durations and other relevant acoustic features will be contextually appropriate and natural by default.

The remaining challenge is to label large corpora of real speech according to a small but sufficiently descriptive set of features so that more of the relevant variations can be indexed and retrieved. This task reduces to a definition of the *perceptually salient* characteristics of speech and of the higher-level factors that contribute to their variation.

Finally, much of the previous research on speech synthesis has been performed on small computers. If we compare the resources currently available to *e.g.*, image processing with those available for speech processing, we see a tremendous mismatch. I maintain that speech is no less complex than image and that if we are to model it accurately, then we need to devote much more processing power and disk space than we are currently considering. In compensation, we see that the currently popular multi-media computing devices are equipped with just such facitities.

6 Acknowledgments

This chapter includes material first presented at the ATR workshop on Computing the Prosody of Spontaneous Speech, and expanded upon in the Symposium on Speaking Styles at the XIIIth Congress of the Phonetic Sciences in Stockholm. I am grateful to colleagues and reviewers for their helpful suggestions and comments.

7 References

[Tal94] Talkin, D., & Wightman, C. W., (1994) "The Aligner: text-to-speech alignment using Markov models and a pronunciation dictionary". In *Proc. ESCA Workshop on Speech Synthesis*, Mohonk, NY. pp 89-92.

[Wig95] Wightman, C., W., & Campbell, W., N., (1995) "Improved labelling of prosodic structures", ATR Technical Report TR-I-053 1994.

[Koh94] Kohler, K., Lex, G., Paetzold, M., Scheffers, M., Simpson, A., Thon, W., (1994) "Handbuch zur Dataneufnahme und Transliteration." in TP14 von Verbmobil - 3.0 Technisches Dokument Nr 11.

[Bru96] Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D., and Touati, P. "On the analysis of prosody in interaction" This volume.

[Cam92] W. N. Campbell & Yoshinori Sagisaka: "Automatic Annotation of Speech Corpora", Proc SST92 Queensland, Australia.

[Kie95] Kiessling, A., Kompe, R., Niemann, H., Noth, E., & Batlinger, A. "Detection of phrase boundaries and accents" *Progress and Prospects of Speech Research and Technology: Proc of the CRIM/ORWISS workshop*, infix, pp266-269, Sankt Augustin, 1995.

[Bla95] Blaauw, E., "On the Perceptual Classification of Spontaneous and Read Speech" PhD Thesis, OTS Dissertation Series, Utrecht University. 1995

[Lin90] Lindblom, B. E. F. (1990) "Explaining phonetic variation: A sketch of the H&H theory". **Speech Production and Speech Modelling** edited by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp 403-409.

[Koh95] Kohler, K, (1995) "Articulatory reduction in different speaking styles". In *Symposium on speaking styles, Proc ICPhS 95*, Stockholm, Sweden.

[Koh96] Kohler, K, "Modelling prosody in Spontaneous Speech " This volume.

[Col92] Coleman, J. C., (1992) "The phonetic interpretation of headed phonological structures containing overlapping constituents". *Phonetics Yearbook 9*, pp 1-44.

[Wha90] Whalen, D., "Coarticulation is largely planned" Journal of Phonetics **18** 3-35. 1990.

[HTK93] Entropic Research Laboratory, Inc, (1993) *HTK - Hidden Markov Model Toolkit* 600 Pennsylvania Avenue, Washington DC 20003.

[Sil92] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., (1992) "ToBI: a standard for labelling English prosody". In *Proceedings of ICSLP92*, volume 2, pp 867–870.

[Cam93] Campbell, W.N. (1993) "Predicting segmental durations for accommodation within a syllable-level timing framework", *Proc Eurospeech-93,*, Berlin, Germany pp 1081-1084.

[Jon95] de Jong, K., (1995) "The supraglottal articulation of prominence in English: linguistic stress as localised hyper-articulation". in *Journal of the Acoustical Society of America 97(1)*, pp 491-504.

[Cam92] Campbell, W.N. (1992) "Prosodic encoding of English speech", *Proc ICSLP-92,*, Banff, Canada pp 663-666.

[Pie92] Pierrehumbert, J. & Talkin, D. (1992) "Lenition of /h/ and glottal stop". In **Papers in Laboratory Phonology II,** eds. G. J. Docherty & D. R. Ladd, Cambridge University Press.

[Gau89] Gauffin, J. & Sundberg, J. (1989) "Spectral correlates of glottal voice source waveform characteristics", **JSHR 32**, pp 556-565.

[Slu93] Sluijter, A., & van Heuven, V. J., (1993) "Perceptual cues of linguistic stress: intensity revisited", *Proc. ESCA workshop on Prosody,*,Lund University, Sweden. pp 246-249.

[Cam95] Campbell, W. N., & Beckman, M. (1995) "Stress, Loudness, and Spectral Tilt", *Proc Acoustical Soc. Japan,* Spring meeting, 3-4-3.

[Ost05] Ostendorf, M., Price, P., & Shattuck-Hufnagel, S., (1995) *The Boston University Radio News Corpus*, Report No BCS - 95 001.

[Hir80] Hirst, D., (1980) "Automatic modelling of fundamental frequency using a quadratic spline function" In *Travaux de l'Institut de Phonétique 15,* Aix en Provence, pp 71-85.

[Cam94] Campbell, W. N., & Black, A. W., (1994) "Prosody and the selection of source units for concatenative synthesis". In *Proc. ESCA Workshop on Speech Synthesis*, Mohonk, NY.

[Cam92] W. N. Campbell: "Synthesis Units for Natural English Speech" SP 91-129, pp 55 - 62.1992.

[Cam93] Campbell, W. N., "Automatic detection of prosodic boundaries in speech". *Speech Communication 13*, pp 343-354. 1993.

[Bla95] Black, A. W., & Campbell, W. N., (1995) "Predicting the intonation of discourse segments from examples in dialogue speech". In *Proc. ESCA Workshop on Spoken Dialogue,* Hanstholm, Denmark.

[Bar95] Barry, W. J., (1995) "Phonetics and phonology in speaking styles". In *Symposium on speaking styles, Proc ICPhS 95,* Stockholm, Sweden.

[Mou93] Moulines, E. & Charpentier, F., (1993)¡ "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". *Speech Communication, Vol 9, nos 5/6,* pp 453-467.

[Fai94] Fais, L., (1994) "Conversation as collaboration: some syntactic evidence", *Speech Communication 15,* pp 230-242.

[Cam95] Campbell, W.N. (1995) "Loudness, spectral tilt, and perceived prominence in dialogues", In *Proc ICPhS 95,* Stockholm, Sweden.

[Hir92] Hirschberg, J., "Using discourse content to guide pitch accent decisions in synthetic speech'. In G. Bailly and C. Benoit, ed, *Talking Machines,* pp 367–376. North-Holland, 1992.

[Hir95] Hirschberg J., (1995) "Acoustic and prosodic cues to speaking style in spontaneous and read speech". In *Symposium on speaking styles, Proc ICPhS,* Stockholm, Sweden.

[Str94] Stenström, A., (1994) *An Introduction to Spoken Interaction.* Longman, London.

[Nak93] Nakatani, C., & Shriberg, L., (1993) "Draft proposal for labelling disfluencies in ToBI". paper presented at 3rd ToBI labelling workshop, Ohio.

[Shr95] Shriberg, L., "Preliminaries to a theory of disfluencies" PhD Thesis, University of California at Berkeley, 1994.

TABLE 1.1. Mean Euclidean cepstral difference for different selection methods

selection based on equal weights	1.9349
selection using weighted features	1.6700
theoretical minimum	1.5456

Notes: (a) 'equal weights' is equivalent to selection using only phonemic environment and provides a measure of the dispersion in the spectra of phonemically identical units in the corpus. (b) 'weighted features' shows the reduction in this distortion that can be achieved by including prosodic descriptors in the selection. (c) the 'theoretical minimum' is defined by selection based on cepstral targets which are impossible to predict in synthesis but allow us to determine the optimal sequence of available units in a given corpus.

TABLE 1.2. Prominence detection

	A	B	C
dur'n & energy:	92%	78%	72%

Showing percentage correct detection of prominence using normalised measures of duration and energy.
Key: A: read grouped, B: read in randomised order, C: interactive

---

TABLE 1.3. ± prominent spectral tilt

	student's t	df
read grouped	35.63	7676
read randomised	19.01	6110
interactive	42.76	6974

Showing the separation in mean spectral tilt between prominent and non-prominent syllable peaks.