FROM READ SPEECH TO REAL SPEECH

W. N. Campbell

ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

ABSTRACT

This paper describes differences in speaking style between read and spontaneous speech from the viewpoint of synthesis research and discusses the development of a set of labels to encode prosodic and segmental variation. Spontaneous speech confronts us with phenomena that were not encountered in corpora of prepared or read speech, and to label them we are increasingly having to identify higher-level units of discourse structure and speaker involvement.

INTRODUCTION

The relationship between speech synthesis and phonetic research is an old one, but we have yet to hear synthetic speech that sounds natural. Some isolated vowels and consonants can be very well replicated and, with careful hand-tuning, even whole utterances can be mimicked, but I am aware of no speech-synthesis-by-rule system that I could yet mistake for a human speaking. Perhaps the best rule-generated synthetic speech that can be heard today is concatenative, using small segments from recorded sequences of real speech and joining them to form novel utterances, but even then the resulting speech loses much of its original naturalness.

I maintain that the reasons for this loss of naturalness are two-fold: a) that typically only a limited number of speech tokens are used to generate a variety of speech, so degradation results from the signal processing required to put them together and modify their prosody, and b) from constraints in the recording of the original speech sequences themselves. Almost all sequences for concatenation are taken from recordings of carefully read speech, and although they may be *phonemically representative*, they are *prosodically constrained* and invariant. Thus, what they encode perhaps models the relevant configurations of the vocal tract for a given speech sequence but fails to adequately model the dynamic articulatory characteristics of the speech.

This paper, focuses not on the *modelling* of speech but on its *characterisation* (or labelling) instead. Using examples illustrating durational characteristics, it shows some effects of the differences in speaking style, and attempts to address the phonetics of spontaneous speech. In doing so, it describes the development of a small set of features sufficient for the description of natural speech such that its variety can be emulated in a synthesis system.

Natural speech

Scientific analysis requires controls, but as Barry has pointed out [1], in the acquisition of speech recordings, these are too often controls on production, with not enough concern for communicative effect. In its natural form, speech is inter-personal and often functionally goal-directed, but in recordings of lab speech (or of speech units for synthesis), where the listener is replaced by a microphone, the speech becomes production-based rather than listener-oriented. As a consequence, the materials we collect and analyse may not be representative of what people do when they speak normally.

For the analysis of natural speech, it is necessary to replace production controls with statistical controls, and use these to study instead large representative corpora of spontaneously produced spoken material. Such corpora are now becoming widely available but the tools for their analysis were developed for a more restricted speaking style. To cope with large volumes of speech, the processing must be automatic, requiring a minimum of manual intervention.

LABELLING SPEECH

Kohler [2] (see also Coleman [3]) has shown that although the articulation of a given sequence of words can vary considerably under different speaking styles according to a cognitively-based reduction coefficient that is dependent on speech act type, a linear segmental representation of canonical citation forms can account well for such phonological reorganisation of speech. He 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. Such a canonical representation is easily accessible from a machine-readable pronunciation dictionary. Thus, given the orthographic transcription of a speech corpus, segmental labelling can be automated to a large extent by using speech recognition technology.

Segmental labelling

By training hidden Markov models (HMMs) corresponding to the phonetic labels in a machine-readable pronunciation dictionary, and generating networks of possible pronunciations for each word, we can use Baum-Welsh re-estimation [4] to model the HMMs closely on each corpus, using the orthographic transcriptions to constrain the alignments, and thereby achieve segmentation accuracy comparable to human transcription [5]. Separate lexical sub-entries are included for some particularly different pronunciation variants such as 'gonna' for 'going to'.

What can be predicted does not need to be labelled. Since the articulation varies according to speaking style, it is sufficient to model the speaking style (or its prosodic correlates) in order to be able to predict the reduction coefficient. The canonical segmental labels allow access to phone-sized portions of the speech waveform from which we can extract prosodic information in order to account for the finer articulatory differences and thus enable us to encode phonation-style characteristics without the need for marking them explicitly.

Materials

The materials referred to in this paper come from four corpora. The first contains readings of 5000 citation-form English words, a subset of these words read one at a time in the form of 200 meaningful sentences, the same sentences read continuously, and 20 minutes of spontaneous interactive monologue (*i.e.*, dialogue with a passive partner). These are in British English from a young adult female speaker, and represent an extreme range of production variation.

The second corpus contains forty-five minutes of professionally read American radio-news speech [6]. It comes from one speaker and exhibits a consistent marked style of production typical of professional announcer speech.

The third, consisting of 300 focus-shifting sentences, produced by an American speaker, illustrates contrastive focus. A set of 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, in the interactive discourse, by (deliberate) misinterpretations on the listener's part.

Finally, from a speaker of American English, is one side of a series of twenty task-related dialogues, performed in a multi-modal environment, alternatively with and without a view of the interlocutor's face [7]. These allow us to compare the speech of one individual, in a highly restricted domain, under a variety of interaction styles.

These corpora were variously labelled at different sites using different transcription conventions, and to compare them it was necessary to relabel all to a uniform style. To do this economically requires definition of a small set of labels that suffice for the complete characterisation of their perceptually salient characteristics. Needless to say, this work is ongoing.

Prosodic labelling

Traditionally, speech has been labelled separately for prosodic and segmental characteristics, but while these features are independently variable, the interaction between them is strong. Segments vary consistently in relation to the prosodic environment so that this segmental variation, in conjunction with the prosodic variation, plays a functional role in chunking the speech and signalling prominence relationships. In read speech at least, boundaries and prominences appear to be the most basic elements of prosodic structure. In locating segments relative to these two dimensions, we can predict much about their articulation.

For example, a given speech segment immediately before a prosodic phrase boundary is likely to be very different from an equivalent one immediately after; it may be considerably lengthened, its amplitude low and decaying, and it may exhibit vocal fry. The segment will also be lengthened in a nuclear accented syllable, but there will be a different profile of lengthening [8] [9] and also increases in spectral tilt resulting from changes in vocal effort [10] [11] [12] [13] and in supraglottal phonation arising from local hyperarticulation [6] [14].

The BU Radio News corpus [6] has been prosodically labelled by hand according to the ToBI conventions [16] to differentiate high and low tones at intonational boundaries and on prominent syllables, and to mark the degree of prosodic discontinuity at junctions between each pair of words. Campbell & Black [17] used this 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.

The test was done by iteratively removing sentences from the radio-news corpus and resynthesising them by concatenation of similar segment sequences selected from the remaining utterances according to suitability of their prosodic environment. This test confirmed that much of the spectral variation in the segments was adequately coded [17]. In one salient example, a sequence of segments across a prosodic phrase boundary were resynthesised using tokens selected from pre- and post-pausal locations such that the 'silence' between them also included an appropriate sharp intake of breath, which made the resulting synthesis sound even more 'natural'[1]. When equivalent tokens from the same segmental sequences were selected from less appropriate prosodic environments, the resulting synthetic speech showed considerable degradation.

Using the radio corpus as training data, Wightman & Campbell [18] defined a set of acoustic, lexical, and segmental features derivable from the phone labels, the dictionary, and the speech waveform, that achieved automatic detection of 86% of hand-labelled prominences, 83% of intonation boundaries, and 88% correct estimation of break indices (at ±1). This was trained using a hybrid combination of a tree quantiser with Viterbi post-processing to maximise the output likelihoods, operating directly on the aligner output.

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 rhyme, the maximum pitch target[2], the mean pitch of the word, intensity at the fundamental, and spectral tilt (using a 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 are all derivable directly from the dictionary used in the aligning.

[1] It should be mentioned here though that because of the limited size of this source database, simple concatenation of these selected units produces noisy synthetic speech, and some (distorting) signal processing would still be required to reduce discontinuities between the selected units.

[2] Pitch targets were calculated using Daniel Hirst's quadratic spline smoothing to estimate the underlying contour from the actual f0.[19]

SPONTANEOUS SPEECH

As an illustration of the contrasts between read and spontaneous speech in British English, we can examine durational structuring, as shown in figures 1 – 4, which plot mean segmental duration against the coefficient of variance (*i.e.,* the standard deviation expressed relative to the mean) for each phone class.

We can see from Figures 1 and 2 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 3 shows the opposite to be the case for the same sequence of words in read sentences. Here the variance increases and there is considerable shortening so that segments are no longer so distinct. 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 complete phrases).

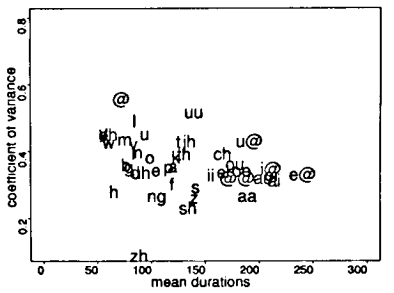

Figure 1: Citation-form words

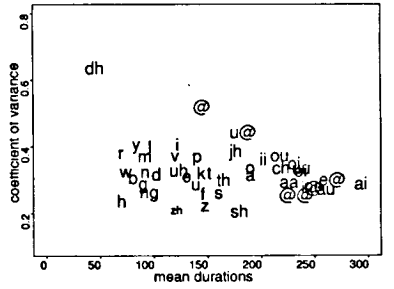

Figure 2: Isolated-word sentences

Rather, the articultion 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 are maximally separated, exaggerating the difference 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 and more varied than in the citation-form readings.

The spontaneous monologue from the same speaker, in Figure 4, shows the same trends exaggerated. we find not only that the mean durations for all segment types are low and uniform,

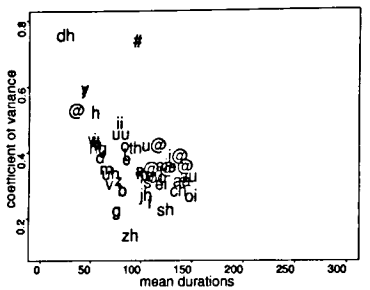

Figure 3: Continuous sentences

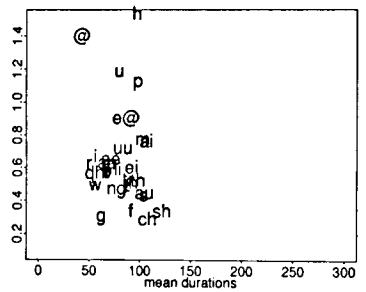

Figure 4: Spontaneous speech

but that the variances are huge - note too that the vertical scale of Figure 4 is *twice* that of the other three figures. When the listener is present, the speaker has a better understanding of the extent of their mutual information, and can rush 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.

Spectral correlates

To show that this difference of style is not unique to durations, nor specific to one speaker, we can compare changes in spectral tilt associated with prominence and focus in read *vs.* interactive speech.

A study of prominence detection, described more fully in [15], 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[3]. The corpus varied prominence in two ways: (1) by asking the speaker to "emphasise" capitalised words in reading, and (2) by eliciting emphatic corrections of feigned misinterpretations.

Table 1: 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

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 that this trade-off is not coincidental, and that

Table 2: ± 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.

the speaker varies her production according to the needs of the discourse context.

Interactive speech

Although much of the labelling of significant levels of information can be done to a large extent automatically for read speech, when extending these techniques to the dialogue corpus and to the spontaneous monologue, we realise the need for extra levels of information to describe the structuring of discourse events and to indicate the switches in speaking style [1] that affect the global prosodic characteristics. Whereas the read speech was highly predictable, the unplanned speech is characterised by bursts of faster and slower sections where the speaker varies her role, and much greater variation in f0 range and pausing as she expresses different degrees of confidence, hesitation, and uncertainty.

Hirschberg has pointed out [20] that the major differences between lab speech and spontaneous speech appear to be prosodic (concerning speaking rate and choice of intonation contour), but there are 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 to have not just the prosodic and segmental information derived from HMM alignment, but also an indication of fluency and 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.

Transcribing the orthography of spontaneous speech requires more than the skills of an audio-typist, and decisions have to be made about marking disfluencies and repairs. To include this information in our labelling, two extra levels of information were added to the ToBI set. One after Nakatani and Shriberg [21], extended the miscellaneous tier of the ToBI transcription to bracket interruptions in the speech flow, and one after Stenström [22]), added IFT (illocutionary force type) speech-act labels. 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, cough, breath, silence.

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 [23] for a discussion), 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 com-

[3] as measured by the relative amount of energy in the top third of an ERB-scaled spectrum (2kHz - 8kHz).

ments 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 first step in this work is to determine the appropriate labels, then we can categorise their prosodic and articulatory correlates. Since human listeners can respond consistently to small speaking-style changes, then the clues must be present somewhere in a higher representation of the signal.

SUMMARY

To summarise the main points of this paper, I have argued that for the efficient characterisation of speech sounds (at least in the context of concatenative speech synthesis), it may not be necessary to label fine articulatory details, nor to attempt a numerical description of the prosodic attributes, but instead to use a higher-level specification of the environment in which they occur.

In read speech, knowing the triphone context of a segment, its position in the syllable, and whether that syllable is prominent, prosodic-phrase-final, or both, allows us to predict much about its lengthening characteristics, its energy profile, its manner of phonation, and whether it will elide, assimilate, or remain robust. Thus for adequate characterisation of speech it is not necessary to label the fine phonetic features explicitly since the higher-level description suffices to include them implicitly.

In the case of *real* speech, however, a significant part of the message lies in the *interpretation* of *how* it was said, and to encode that level of information, we need to incorporate labels for discourse and communication strategy; we need to estimate the state of mind of the speaker, her commitment to the utterance, and the role of that utterance in a greater discourse. This is future work.

Finally, to return to synthesis, if we consider a hugely finite corpus of natural speech as a source of units for concatenative synthesis then, instead of disruptively warping a segment to fit a predicted context, it would be possible to select an appropriate segment from amongst the available variants. Furthermore, if that corpus were adequately labelled in terms of all the contributing factors (i.e., with phonemic, phrasal, prosodic, speech-act etc., labels), then it would no longer even be necessary to predict fine details of the speech at all; it would be enough to select a segment with the same labels to characterise the desired target speech. The durations and other relevant acoustic features would be contextually appropriate and natural by default.

The remaining challenge is to label large corpora of real speech according to a small and sufficiently descriptive set of features so that all the relevant variations can be indexed and retrieved. This reduces to a definition of the *perceptually salient* characteristics of speech, which in turn enables us to use only a large speech corpus instead of a huge one without loss of naturalness.

REFERENCES

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

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

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

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

[5] 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.

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

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

[8] 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.

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

[10] 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.

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

[12] 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,

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

[14] 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.

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

[16] 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.

[17] 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.

[18] Wightman, C., W., & Campbell, W., N., (1995) "Improved labelling of prosodic structures", IEEE Trans. Sp. & Audio, submitted.

[19] 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.

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

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

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

[23] 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.