Non-segmental analysis and synthesis based on a speech database

Andrew Slater and John Coleman

Phonetics Laboratory, University of Oxford,
41 Wellington Square, Oxford OX1 2JF, UK.

Email: andrew.slater@phonetics.oxford.ac.uk, john.coleman@phonetics.oxford.ac.uk

ABSTRACT

This paper reports on experiments in non-segmental speech analysis and synthesis using parameters derived from a speech database of British English monosyllables. The database includes almost every onset, nucleus and coda, and almost all onset-nucleus and nucleus-consonant combinations occurring in English. Acoustic parameters including f0, formant frequencies and bandwidths, and amplitude of voicing were determined for each token in the database. Fine duration differences within minimal pairs are analyzed using dynamic time warping techniques, avoiding the need for manual segmentation. For each parameter, a matrix of distances between all samples of the two words is calculated, together with a minimal path through the matrix (the **warp path**). The set of warp paths for all parameters identifies the nature and location of acoustic differences between the words, including locations of temporal expansion and compression. Preliminary experiments using dynamic time warping for non-segmental synthesis are also discussed.

1. INTRODUCTION

1.1 Formant synthesis and naturalness

Speech synthesis techniques are conventionally divided into two approaches: concatenative synthesis and formant synthesis-by-rule. Concatenative synthesis, being based on recorded speech, has the advantage of sounding more natural, which partly accounts for its popularity in practical applications. However, concatenative techniques typically employ units which are not linguistically motivated, such as diphones or demisyllables. Consequently, concatenative synthesis affords little insight into the linguistic structure of the language being synthesized.

Formant synthesis-by-rule uses units which are linguistically motivated, including features whose domain is not restricted to segment-sized units. However, it sounds less natural, due in part to highly stylised parameter dynamics. Hawkins and Slater [4] claim that the less natural quality of formant synthesis is due to the failure to model aspects of systematic spectral variation that are not critical to phoneme identity. Capturing this type of variation in formant synthesis usually requires extensive handcrafting which is time consuming, brittle, prone to error and requires expert knowledge that is not transparent in the rules or lookup tables.

We employ techniques for automatic extraction of parameters from carefully recorded speech. In this way, we are able to model aspects of the natural variation of real speech using a standard formant synthesizer [12].

1.2 Timing models in text-to-speech

Most TTS systems (e.g. [7]) model speech as a linear sequence of phoneme-sized segments, with parameter targets located relative to the segment boundaries. Duration variability is typically modelled as the interaction of factors such as position in phonological or prosodic structure and overall speech rate. Temporal expansion or compression is treated as uniform within each segment, in spite of evidence from empirical studies (e.g. [2]) that such effects are non-linear.

For example, the initial part of the diphthong /aɪ/ is longer in “bide” than in “bite”, whereas the glide portions of the two words have similar durations. It is difficult to account for this type of segment-internal structure within the standard segmental model. We believe these kinds of difference are critical to producing natural-sounding and intelligible synthetic speech.

There have been some attempts to model segment-internal timing differences, e.g. [3, 5, 6]. These theories describe speech segments as a series of steady states and transitions. Transition regions are relatively stable, whereas steady states are subject to expansion or compression. Hertz [5] and Hertz and Huffman [6] argue that the “phone and transition” model offers simpler accounts of vowel duration and other timing phenomena, such as aspiration. However, a serious limitation of both subsegmental and traditional phoneme-based models is their reliance on an initial manual segmentation of words involved in a particular contrast. This is in practice highly problematic and often subject to ad-hoc criteria. Even where segment boundaries may be reliably identified, different parameters might not be naturally segmented at the same point as other parameters. Subsegmental approaches do not solve the problems of temporal modelling; they merely push them down to a finer level of analysis.

1.3 Dynamic time warping

We employ the technique of **dynamic time warping** (DTW) as a means of deriving segment internal temporal structure without referring directly to segment or subsegment boundaries. DTW was developed in speech recognition [10] as a means of non-linear time-alignment of two signals. A matrix of distances is computed between all pairs of samples of two utterances. An algorithm is used to compute the minimal cost path (the **warp path**) through the distance matrix. Analysis of the shape of the warp path can reveal acoustic-phonetic differences between similar words, such as minimal pairs. Previous studies have used DTW to analyze polysyllabic shortening [8] and durational characteristics of minimal pairs differing in coda voicing [1]. In section 2 we

describe the preparation of a speech database for DTW research. In section 3 we report on our own study of coda voicing, and in section 4 preliminary experiments on the use of DTW for non-segmental synthesis are discussed.

2. PREPARATION OF SPEECH DATABASE

Previous studies (e.g. [8]) have shown DTW to be highly sensitive to the accuracy of the acoustic parameters used by the algorithm. A database containing high quality audio files, together with extracted acoustic parameters was prepared.

2.1 Method

The database comprises 5 tokens of each of 1066 monosyllabic words chosen from the computer-readable Oxford Advanced Learners' Dictionary [9]. The 8444 monosyllabic words in [9] were parsed into the syllable constituents onset, nucleus and coda, and database items were selected to include almost every onset, nucleus and coda, and almost all onset-nucleus and nucleus-consonant combinations. Some categories were excluded due to their marginal status in British English (e.g. onset /ʃm/). The dataset was expanded slightly to include a number of additional minimal pairs differing in coda voicing for the experiment described in Section 3.

Each word was embedded in a carrier phrase according to the phonological category of the initial and final phonemes of the word, as shown in **Table 1**. The sentences were randomised in blocks of 1066 to give 5 tokens of each. They were read by a male speaker of British English, aged 33, in a sound-treated room. Speech was recorded using DAT and digitised at 16kHz using a Silicon Graphics Indy on-board A–D convertor. Extracts of each sentence were taken from the beginning of the /t/-closure in "utter" or "uttered" to the burst of the medial stop in "today" or "again".

Word		**Carrier phrase**
Initial segment	Final segment	
cons	cons	Can you utter — again please?
cons	vowel	Can you utter — today please?
vowel	cons	Have you uttered — again please?
vowel	vowel	Have you uttered — today please?

Table 1: Carrier phrases used for database words.

For each token, several acoustic parameters were extracted. f0, frequencies and bandwidths of F1, F2, F3 and F4, and amplitude of voicing were obtained using the ESPS/xwaves "formant" program. Formant tracks were visually inspected and tracking errors manually corrected using wideband spectrograms and LPC spectra. Amplitude of friction was estimated from the amplitude integral of portions of the linear prediction error signal corresponding to unvoiced speech. Acoustic parameters were recorded at 5ms frames, which is adequate for formant synthesis.

2.2 Discussion

A phonologically rich acoustic database of British English monosyllables has been collected. A basic set of acoustic parameters has been extracted for each token in the database. Impressionistically, resynthesis using the extracted parameters as input to the Klatt synthesizer [12] yields speech which is not as good as the original recordings, but which is more natural in some respects than synthesis-by-rule output. It is hoped that the database will be of more general use than the DTW research reported here, and it is our intention to release the database and associated parameter files to the speech community.

3. STUDY OF POST-VOCALIC VOICING

It is well established (e.g. [11]) that English vowels are longer before voiced obstruents than before voiceless obstruents. This is usually modelled in TTS using context-sensitive vowel lengthening or shortening rules. This experiment examines the temporal details of minimal pairs whose members differ in postvocalic voicing, using DTW techniques.

Time warps were computed between centroid trajectories of words within each minimal pair, and analysis of these time warps yields details of the temporal differences between the words. Our methodology is based on [1] in many respects, applied to our database of British English. However, in anticipation of using the resultant time warps for formant synthesis, we use the formant synthesis parameters detailed in Section 2.1 rather than cepstrum delta-cepstrum space, as in [1]. We also compute one time warp *per parameter* for each minimal pair.

We expected differences in time warps to be conditioned by vowel identity (diphthong vs. monophthong) and by presence or absence of a following or preceding sonorant.

3.1 Method

48 voicing pairs were selected from the list of 56 pairs in [1]. (Pairs with contrastive vowel phonemes in British English, e.g. **can't-canned**, were excluded.) They are: bait, bayed; bet, bed; belt, belled; bent, bend; bite, bide; bleat, bleed; boot, booed; burnt, burned; caught, cord; cite, side; coat, code; dwelt, dwelled; faint, feigned; felt, felled; girt, gird; gloat, glowed; great, grade; hurt, heard; heat, heed; height, hide; lent, lend; light, lied; mate, made; meant, mend; pat, pad; paint, pained; pant, panned; plate, played; pleat, plead; plot, plod; rate, raid; root, rude; rout, rowed; wrote, rode; set, said; seat, seed; sent, send; shot, shod; shoot, shoed; smelt, smelled; spent, spend; spurt, spurred; suit, sued; tote, toad; trot, trod; weight, weighed; went, wend; wet, wed.

Dynamic time warping was used in two ways in this experiment. First, it was used to construct centroid trajectories from the 5 tokens of each word in the database. This was done to minimize the noise present in individual tokens. Second, DTW was used to

compare paired centroids. In all cases, time warping was performed using the entire database extract, as defined in Section 2.1. We used a slightly modified version of the usual DTW algorithm. The allowable moves for the warp path were (i) 1 frame horizontal (ii) 1 frame vertical or (iii) 1 frame horizontal and 1 frame vertical. The distance measure used was **|log(y/x)|** where samples x and y are single parameter values from a given pair of words. (The reason for this distance metric is discussed in Section 4.) A lookahead of 1 frame was employed, and diagonal moves were weighted favourably. Centroid trajectories were computed for each word, for each parameter separately, as follows: (i) Pairwise time warps of the 5 recorded trajectories were computed, giving 25 new trajectories. (ii) A set of 5 new trajectories — "candidate centroids" — were computed from the medians of the trajectories computed in (i). The candidate centroids nearly converge after 2 iterations. When candidate centroids are computed for all 10 acoustic features, a single centroid is chosen from each set of 5 candidates, using the lowest *median sum distance* of the candidates from the original trajectories. This choice is cast as a vote for a set of 10 related parameter centroids of the same length. Thus, for each word in the dataset, we have 10 centroid trajectories, one per acoustic feature. The quality of the centroids was confirmed by listening to resynthesis of each word using the centroids.

DTW was then used to compute one time warp per parameter for the centroid trajectories of each minimal pair. (In each case, the voiceless member was treated as the reference.) The sonorant portion of each centroid (i.e. the vowel plus any tautosyllabic sonorant consonants) was manually delimited using waveforms and spectrograms. A measure of *maximal expansion* was calculated within the sonorant portion of the reference (voiceless) member using the deviation of the time warp from the distance matrix diagonal.

3.3 Results and discussion

Correctness of warp paths

The scheme for weighting moves in warp paths should be such that they do not deviate excessively from the diagonal, yet have sufficient variability in slope to be useful: the shape of the warp path is critical in order to provide useful insights into the temporal details under investigation. However, selection of the "best" weighting scheme is problematic: if the pro-diagonal weighting is insufficient, some expansions may be spurious and the warp path may go awry. On the other hand, excessive pro-diagonal weighting yields warp paths which are unnaturally flat, i.e. showing uniform expansion.

Computation of centroids

DTW proved successful for the computation of centroid trajectories. Figure 1 shows an example of 5 recorded trajectories and the 5 candidate centroids for F1 of **height**. In general, words resynthesized from the centroids sound no worse than, and in some cases better than, resynthesis of the original tokens.

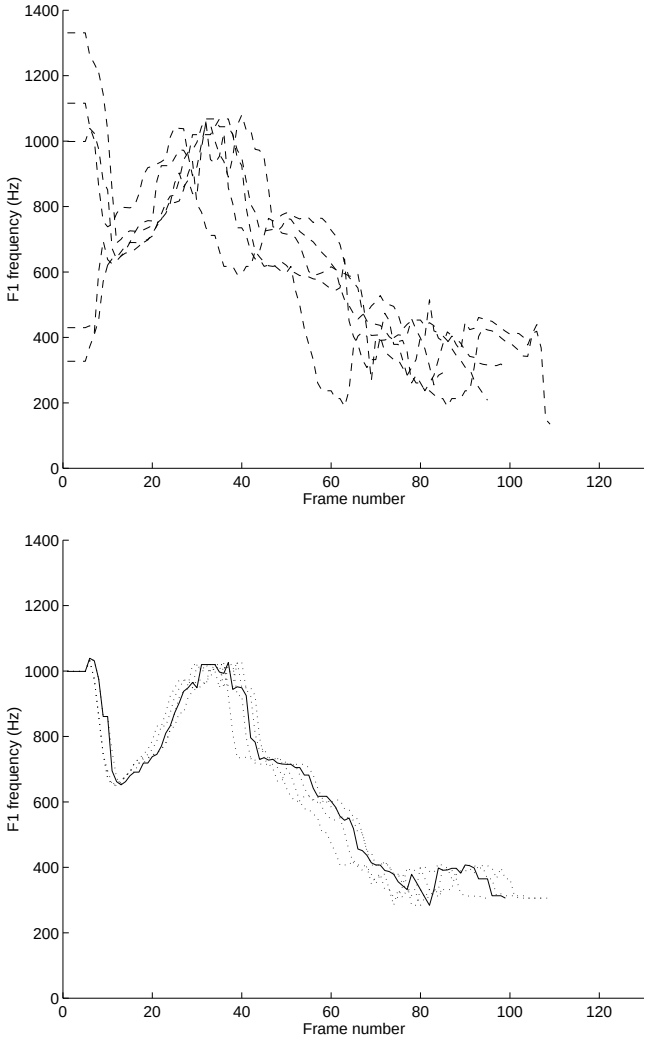

Figure 1: F1 parameter values for **height**. (Upper panel) recorded trajectories; (lower panel) candidate centroids, solid = chosen centroid.

Warping of centroids

Figure 2 shows an example of a time warp distance matrix, and the computed warp path through the matrix, of F2 for the minimal pair **great-grade**.

Location of expansion

Location of maximal expansion proved to be highly variable: in some pairs (e.g. **bite-bide**, **height-hide**, **great-grade**) maximal expansion is found early in the sonorant portion of the syllable, in some cases even in sonorant onset consonants. In other cases, maximal expansion occurred in the vowel, either towards the beginning or the end, and in some cases including postvocalic sonorant consonants (e.g. **meant-mend**, **went-wend**). However, we can make no simple phonological generalisations, and the location of maximal expansion varied from one parameter to

another. An SPSS general linear model of maximal expansion was constructed for each parameter, in which the dependent variable, location of expansion, was expressed as a percentage of the duration of the sonorant. Three factors were examined: identity of prevocalic sonorant (5 levels), identity of vowel or diphthong (11 levels) and identity of following sonorant (3 levels). The only significant factor for any of the 10 parameters was preceding sonorant ($df = 4$, $p < .05$) for F1.

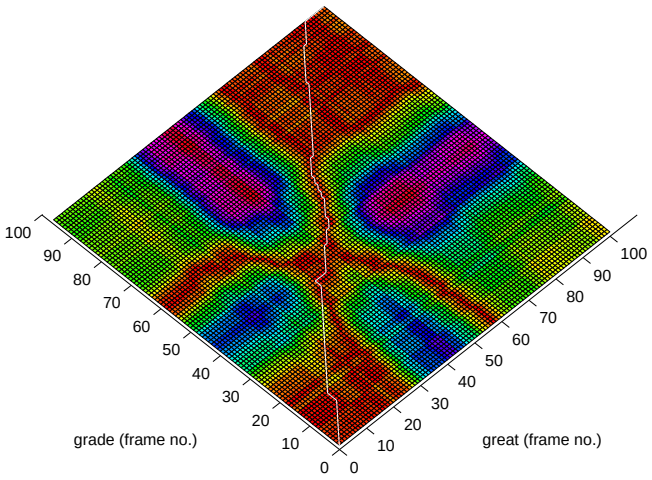

Figure 2: Distance matrix and warp path for F2 of **great-grade**. Greyscale represents distance (dark = small; light = large). The white line is the warp path.

4. EXPERIMENTS IN NON-SEGMENTAL SYNTHESIS

Since the set of warp paths quantifies the differences between word pairs, the warp paths can also be used to transform one utterance into another for the purpose of synthesis. We compute a time warp for synthesis purposes as follows. A distance matrix for each parameter P of words x and y is calculated as before. The DTW produces a sparse warp matrix W. Cells in W are 0 except for a single entry in each column. Non-zero cells $W_{i,j}$ have value Py_j/Px_i where i and j are frame number of x and y associated by DTW, and Py_j/Px_i is the ratio of the values of parameter P for each word for those frames. We can convert parameter x into parameter y using matrix multiplication, since y=xW. We call this operation **warp path transformation**. This is repeated for each parameter in turn, to generate a complete set of resynthesis parameters. Parameter files and the speech they generate resynthesized using this method are always perfect copies of the originals, as the warp path is just a sample-by-sample ratio of the two signals. Constraints on the DTW algorithm which are critical for speech analysis are irrelevant for analysis-resynthesis.

Analysis-resynthesis requires use of the parameters extracted from the target word in calculating the warp path used in the resynthesis phase. We are currently investigating the possibility of generalising the technique for the synthesis of novel utterances. This research employs clusters of phonologically related warp matrices in order to use transformations computed for a phonological relationship in one set of examples (e.g. **bite → bit,** or **my → mine**) to model parallel cases (e.g. **light → lit, pie → pine**).

6. REFERENCES

1. van Santen, J., Coleman, J.S. and Randolph, M. "Effects of postvocalic voicing on the time course of vowels and diphthongs", [Abstract] *J. Acoustic. Soc. Amer.,* 92(4*)*, Part 2: 2444, 1992, and unpublished notes.

2. de Jong, K. "An articulatory study of consonant-induced vowel duration changes in English", *Phonetica*, 45: 156–174, 1988.

3. Gay, T. "Effect of speaking rate on diphthong formant movements", *J. Acoustic. Soc. Amer*., 44: 1570–1573, 1968.

4. Hawkins, S., and Slater, A. "Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech.", *Proc. ICSLP 94*, 57–60, 1994.

5. Hertz, S.R. "Streams, phones and transitions: toward a new phonological and phonetic model of formant timing", *Journal of Phonetics* 19, 91–109, 1991.

6. Hertz, S.R. and Huffman, M. "A nucleus-based timing model applied to multi-dialect speech synthesis by rule", *Proc. ICSLP 92*, 1171–1174, 1992.

7. Allen J., Hunnicutt M.S. and Klatt D. *From text to speech: The MITalk system*, 1987.

8. Macchi, M.J., Spiegel, M.F. and Wallace, K.L. "Modeling duration adjustment with dynamic time warping", *Proc. ICASSP 90*, 333–336, 1990.

9. Mitton, R. A computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English. *ftp://ota.ox.ac.uk/pub/ota/public/dicts/710/text710.dat*, 1992.

10. Sakoe H. and Chiba S. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-26, 43–49, 1978.

11. Peterson, G.E. and Lehiste I. "Duration of syllabic nuclei in English", J. Acoustic. Soc. Amer., 32: 693–703, 1960.

12. Klatt, D.H. "Software for a cascade/parallel formant synthesizer", J. Acoustic. Soc. Amer., 67: 971–995, 1980.