Proceedings of the Institute of Acoustics

CONTEXT-SENSITIVE DIPHONES AS UNITS IN SPEECH SYNTHESIS*

Jo W.M. Verhoeven

University of Edinburgh, Centre for Speech Technology Research, 80 South Bridge, Edinburgh EH1 1HN

1. INTRODUCTION

The text-to-speech system for British-English that is presently being developed at the Centre for Speech Technology Research uses diphones as its basic units in speech synthesis. A diphone is a portion of speech stretching from the midpoint of the spectrally steady-state part of one sound to the midpoint of the steady-state portion of the next sound. Consequently, diphones contain the transitions between segments. These diphones are extracted from natural speech recorded in carefully controlled conditions and are stored in a diphone inventory. In order to synthesize speech, the diphones required to build an utterance are selected from this inventory and strung together in the appropriate order. At first sight, the clear advantage of this method is that the synthesis problem is simplified to joining steady-state speech portions, while the complex transitions between sounds are available from the diphone inventory.

The basic expectation of diphone synthesis, i.e. that diphones will join smoothly together has, however, to be regarded as overoptimistic. Intrinsic phonetic variability and slight variations in speaking rate cause differences in the realization of the articulatory targets for each phoneme, which is necessarily reflected in the spectral characteristics of the diphones. As a result, spectral discontinuities between abutting diphones in speech synthesis are unavoidable. In addition, variations in recording conditions can cause discontinuities in amplitude at diphone joins. Both factors can potentially degrade the speech quality by introducing audible clicks and a general impression of roughness (Isard & Miller [1]).

In order to minimize the amount of smoothing that is required to overcome these problems, it was decided to direct research effort towards the development of diphone segmentation methods, which a priori guarantee minimal spectral discontinuity between all theoretically possible pairs of diphones. Before discussing these methods, the traditional method of hand-segmentation will be described.

* This work was supported by the Information Engineering Directorate/Science and Engineering Research Council as part of the IED/SERC Large Scale Integrated Speech Technology Demonstrator Project (SERC grants D/29611, D/29628, D/29628, F/10309, F/10316, F/70471) in collaboration with Marconi Speech and Information Systems and Loughborough University of Technology.

2. HAND-SEGMENTATION

The point of departure of this investigation was a diphone inventory extracted from 2169 nonsense words, which were read by a male native speaker of British-English. The structure of these words is described in detail in Isard & Miller [1]. The main aim of embedding the diphones in nonsense words, rather than extracting them from entirely natural speech, is to execute careful control over the phonetic environment in which the diphones occur. Since all the diphones, except those involving syllabic /l/, /m/, /n/ and the neutral vowel schwa, occur in stressed syllables, preceded and followed by a phonetically neutral context, it is possible to achieve a high degree of consistency in pronunciation.

The nonsense words were segmented by means of an automatic segmentation algorithm, which is described in Taylor [2]. Subsequently, the diphone boundaries were determined by a trained phonetician, who had the appropriate tools at her disposal to display the time-aligned waveforms and spectrograms of the nonsense words together with the segment boundaries obtained by the segmentation algorithm. It was possible at any time to play back portions of the speech signal. In order to increase the consistency of the segmentations, the phonetician was guided by the following set of simple segmentation criteria:

(a) Vowels, nasals and liquids are segmented in the middle of their spectrally steady-state portion.
(b) Fricatives are segmented in the middle.
(c) Stops are cut in the middle of their silent component.
(d) Diphthongs are segmented in the middle of their first steady-state region.

Although most of these criteria are relatively straightforward, (d) can be expected to cause major problems, in that the first steady-state portion cannot always be adequately identified. In most cases, these steady-state portions are extremely short or absent altogether. Hence, diphthongs potentially constitute a class of sounds where segmentation inconsistencies can be expected.

This method has resulted in a complete diphone inventory of British-English, the first four entries of which are given as an example in table 1:

r-uh	d001.vox	552	615	649
r-e	d002.vox	637	705	756
r-i	d003.vox	594	658	692
r-o	d004.vox	518	581	628

Table 1 : First four diphone entries from the CSTR diphone inventory for British-English.

Besides the identity of the diphones and the speech files from which they have been extracted, the inventory contains timing information (in msec) about three boundaries: the starting point of the diphone, the segment boundary and the endpoint of the diphone. The segment boundary

provides a fixed reference for the calculation of segment durations in the synthesis process. The complete inventory consists of 2390 diphone entries.

Once the diphone inventory was established, the total spectral discontinuity in synthesized speech on the basis of these hand-segmented diphones was assessed. For this purpose, all the diphones ending in a particular sound were joined with all the diphones beginning in this sound: i.e. all the diphones ending in [a] were combined with all the diphones beginning with [a] etc. Thus, all theoretically possible diphone combinations were obtained. Subsequently, a linear distance measure was calculated between adjoining frames of cepstral coefficients in each diphone combination to give an indication of the spectral mismatch at the diphone joins. Finally, the measure of spectral discontinuity (MSD) was averaged for the different sound classes (figure 1). It should be indicated that, although these measures are not very informative in themselves, they are essential as reference values in the assessment of the other segmentation techniques that will be discussed below.

3. VECTOR CLASSIFICATION

In the hand-segmentation method, diphone extraction is based on the segmentor's judgement of the spectral characteristics of each individual diphone in isolation, in that the characteristics of other diphones in the inventory are not taken into account. An alternative approach is to determine the boundaries of a diphone in the inventory in such a way that they yield minimal spectral discontinuity with respect to all the other diphones they can be combined with in the actual synthesis of speech. In order to achieve this, the spectral composition of a diphone has to be evaluated with respect to all these other diphones, before a decision can be made about its ideal boundaries. This constitutes the essence of the vector classification method.

In vector classification, the frames of cepstral coefficients in all the instances of a particular sound in the inventory are submitted to a classification algorithm, which determines a mean vector of cepstral coefficients or a mean cepstrum. The ideal diphone boundary can subsequently be taken as the vector of coefficients which constitutes the closest match to the mean cepstrum. More concretely: in all diphones involving [i], the phoneme boundaries of [i] are available from the automatic segmentation process. Subsequently, all the frames of cepstral coefficients of every occurrence of [i] are extracted and used to calculate a mean cepstrum. Subsequently, all the frames of cepstral coefficients of each [i]-diphone are compared to the mean cepstrum and each [i]-diphone is segmented at a point where its cepstrum is closest to the mean cepstrum. This procedure was used to segment vocalic regions, since such sophistication is wasted on silent regions of for instance stops and on fricatives, which are characterized by fairly large frame-to-frame variations.

In order to evaluate whether this segmentation method reduces average spectral mismatch, the total amount of spectral discontinuity resulting from this method was calculated and compared with that arising from the hand-segmented diphone inventory described above. A comparison of these average measures for both segmentation methods is given in figure 1:

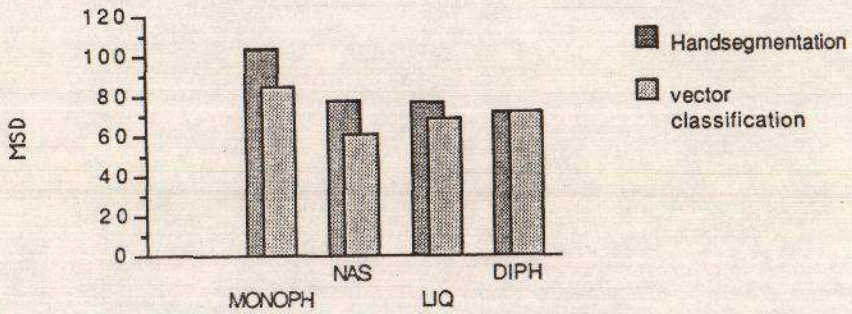

Figure 1: Average measure of spectral discontinuity between all theoretically possible diphone combinations involving monophthongs, nasals, liquids and diphthongs.

The comparison of the two methods indicates that vector classification performs slightly better than the hand-segmented method, in that an overall reduction in mismatch of 11% is obtained.

At first sight, this result seems to suggest that the vector classification method of determining diphone boundaries yields a significant reduction in spectral discontinuity in the diphones. The improvement is however more apparent than real. Closer inspection of the MSD for individual diphone combinations reveals that in roughly half the cases, the vector classification method in fact increases spectral discontinuity. In the remaining half, the method reduces discontinuity in such a way that the average gains slightly outperform the losses. As a result, it has to be concluded that a diphone inventory extracted by vector classification alone cannot be considered an improvement in real terms, since in about 50% of the cases, speech output is likely to be even more degraded than by using the hand-segmented dictionary. In this output quality perspective, the losses outperform the gains. A possible solution to the problem is the compilation of a hybrid diphone inventory, which contains the best-performing diphones from both segmentation methods. This would necessarily lead to an overall reduction of spectral mismatch of well over 11%.

4. CONTEXT-SENSITIVE SEGMENTATION

The essence of both methods described so far is that each diphone in the inventory has a single set of fixed boundaries. A possible alternative is the use of diphone-like units with variable boundaries, which can be called context-sensitive diphones. These are essentially flexible units: their boundaries are not determined in isolation, but are made dependent on the left- and right-hand diphones they combine with in a concrete utterance. The segmentation criterion for each diphone is such as to minimize spectral discontinuity between the diphone in an utterance and its left- and right-hand neighbour.

CONTEXT-SENSITIVE DIPHONES

In practice, the starting point for diphone extraction is again an inventory of phoneme combinations, the boundaries of which have been established by the automatic segmentation procedure. At the point in the synthesis stage when it is known which diphones are required to produce an utterance, full phoneme combinations from which the diphones are to be extracted are selected from the inventory. Subsequently, all combinations of frames in adjoining phonemes are compared and a measure of spectral discontinuity is calculated. The ideal boundaries are then chosen as those frame combinations which yield the lowest mismatch.

So far, two versions of this technique have been implemented. In the first, the algorithm searching for the ideal frame combination is not restricted in time, in the sense that all the frames of abutting phonemes are compared with each other. In the second version, the algorithm is restricted in its search to the middle third portion of abutting phonemes. The latter method only has been fully developed and tested, since it is expected to do more justice to phonetic reality, in that context effects of neighbouring sounds are likely to have disappeared in the middle portion of phonemes.

As for the method of vector classification, the average amount of spectral discontinuity obtained by the use of context-sensitive diphones was calculated and compared to the other methods. It is summarized in figure 2:

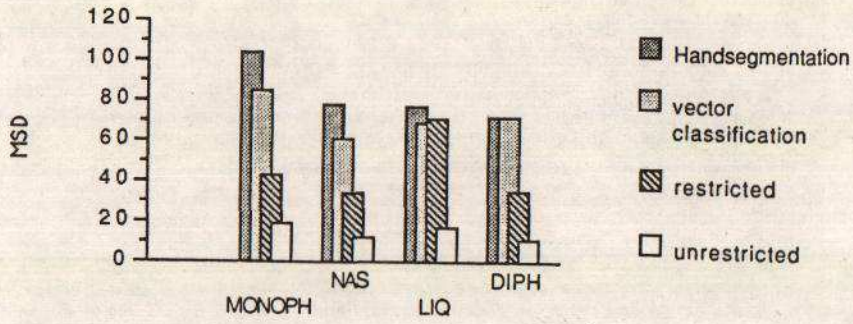

Figure 2: Comparison of average spectral discontinuity in the four segmentation methods for monophthongs, nasals, liquids and diphthongs.

It can be seen from this comparison that synthesis by means of hand-segmented diphones produces the largest amount of spectral discontinuity. Synthesis by means of restricted context-sensitive diphones reduces the spectral discontinuity by 55% on average. Although the gain is quite outspoken for monophthongs, nasals and diphthongs, it is only marginal for liquids. Finally, unrestricted context-sensitive diphones reduce the average spectral mismatch by 81% as compared to the hand-segmented diphones. This can be regarded as the maximal reduction of spectral mismatch attainable for this inventory.

5. DISCUSSION

The diphone segmentation techniques proposed above provide a principled basis for diphone extraction, in that they explicitly take into account the spectral characteristics of the other sounds with which each diphone may combine in speech synthesis. They abandon the principle of cutting diphones in isolation and hence are able to reduce the total amount of spectral mismatch in speech synthesis. A statistical evaluation of these techniques shows that the vector classification method performs slightly better than hand-segmentation. Context-sensitive diphone extraction can be regarded as a superior alternative, in that a reduction of 81% can be obtained with the unrestricted algorithm. The restricted algorithm yields an average reduction in mismatch of 55%.

Although unrestricted context-sensitive diphones yield the most dramatic reduction in discontinuity, there are good reasons to believe that these units may not be the best option for speech synthesis. Informal experimentation with this technique has shown that in quite a number of instances, the algorithm tends to cut out a sound almost completely, which clearly is to be avoided. This problem cannot occur in restricted context-sensitive diphones, since boundary extraction is limited to a particular area of the diphone. Hence there will always be a long enough stretch of diphone available to synthesize the intended sounds. Therefore, two inventories of restricted context-sensitive diphones have been compiled at CSTR for two male speakers of British-English and a third inventory for a female speaker is in preparation.

It should be pointed out that the context-sensitive diphone segmentation technique can operate in two different ways. On the one hand, the diphones can be extracted on-line, when they are actually needed in speech synthesis. This has the advantage that no extra storage space is required. On the other hand, the diphones can be segmented off-line, which involves the storage of a large table of all hypothetically possible diphone combinations with their associated boundaries. In synthesis, the required diphone combinations are selected from this table and strung together in the appropriate order.

Informal listening tests with speech researchers familiar with synthetic speech have shown that the use of context-sensitive diphones lead to a considerable improvement of the speech output quality. A more systematic perceptual evaluation of this technique is presently being carried out. Besides this perceptual evaluation, there are a number of aspects which still have to be further investigated. Firstly, the possibility of weighting the cepstral coefficients has to be considered. Presently, a linear distance measure on unweighted cepstral coefficients is used. It has already been informally observed that this sometimes results in a better match of the higher formants, whereas the mismatch between the lower formants at diphone joins increases. This is clearly undesirable and the use of mel-scaled cepstral coefficients as a basis of the distance measure may solve this problem. Research into this aspect is already underway. Secondly, it is to be investigated whether this diphone segmentation technique can be combined meaningfully with durational modelling in the text-to-speech system, by choosing the diphone boundaries in such a way that the appropriate duration of the sounds are obtained while at the same time minimizing spectral mismatch.

REFERENCES

[1] S. Isard & D. A. Miller, 'Diphone synthesis techniques', IEE Conference Publication no 258 p77-82 (1986)
[2] P. Taylor, 'Automatic diphone segmentation using HMM', 3rd Australian International Conference on Speech Science and Technology, in press (1990)

ISBN 1 873082 12 6 pp.1 - 282
ISBN 1 873082 14 2 set of two