Heading for a diphone speech synthesis system for Dutch

B.A.G. Elsendoorn

Abstract

A survey is given of the progress made in the development of a diphone speech synthesis system for Dutch. Some programs that facilitate concatenation and input of intonation contours are briefly described and the use of 'special' diphones is explained.

Introduction

An earlier report (Elsendoorn & 't Hart, 1982) showed a fairly promising start in the development of a diphone-based speech synthesis system for Dutch. The very simple form of diphone concatenation that was used in the early system resulted in remarkably acceptable speech. However, the system needed improvements and progress was not always without difficulty. Some of the developments are described below.

Input

Two computer programs have been devised by Niesen and De Jong to facilitate diphone concatenation. The first program takes as its input a pseudo-phonetic string of characters, which is converted into a diphone string. An example of such strings is given in the top portion of Figure 1. The program then checks whether all diphones exist in the inventory, in which case the diphones are concatenated. In the resulting speech file, the diphones are represented by LPC-coded formant/bandwidth values, amplitude, F_0 and voiced/unvoiced information for every 10 ms of speech.

```
PSEUDO-PHONETIC STRING        :    cn apcl pcr dax     (trl.: an apple a day)

DIPHONE STRING                : #C CN NA AP PC CL LP PC CR RD DA AX X#

DIPHONE STRING (new version) : #C CN N? ?A AP PC CL LP PC CR RD DA AX X#

INTONATION MARKERS            :           //                         \\
```

Fig. 1. An example of the input strings used in the two programs.

This speech file can then be fed into the second program, by means of which the user can add intonation contours. The user is provided with the same diphone string, below which he/she can specify prominence-lending rises or falls or more gradual pitch movements. This is illustrated on the fourth line of Figure 1. These movements are superimposed upon a declination line that is automatically generated by the program. The movements have the standard values for Dutch, viz. a 120-ms duration and a 6-semitones excursion. In normal speech the correct starting point for pitch movements is related to the vowel onset ('t Hart & Cohen, 1973). Therefore phoneme boundaries had to be marked in all diphones to determine the correct starting point for pitch movements.

The diphone-concatenated speech can also be subjected to a smoothing algorithm written by the author. This algorithm interpolates amplitude, formant and bandwidth values across diphone boundaries. Markers have been introduced into all diphones to indicate where the interpolation must start or end. The number of 10 ms frames across which interpolation takes place varies according to the type of phoneme (20 ms for short vowels to about 80 ms for long vowels and voiceless fricatives). Obviously plosives are not subjected to smoothing.

More diphones

The diphone inventory has gradually expanded to more than 1850. Those diphones that have been added include 1) consonant-glottal stop diphones and glottal stop-vowel diphones (represented below as C? and ?V, respectively), 2) vowel-vowel diphones (VV), 3) consonant-/e:/ and consonant-/o:/ diphones for /e:/ and /o:/ followed by /r/, 4) consonant-/h/-vowel triphones and 5) unstressed /ə/ diphones.

1) In the early system, CV diphones had been truncated from the second syllable of nonsense words of the CəCVCə type, in which all consonants were identical. This was done to obviate coarticulation effects that might have occurred if existing words had been used. However, C? and ?V diphones have been added to account for glottal stops that may occur at word boundaries, as in 'een appel' (/ən?apəl/). In this case, the use of the 'normal' /na/ diphone (truncated from /nənanə/ without a glottal stop) did not lead to natural speech (/ənapəl/, cf. the second line of Fig.1). Concatenation of utterance-final consonants with utterance initial vowels did not lead to acceptable results either, since the amplitude decay time of the C# diphone was too long. Therefore C? diphones were segmented, which display a more abrupt decay in amplitude. Besides, ?V diphones were added, since the amplitude rise time for ?V is somewhat shorter than for #V diphones. The new diphone string is presented in the third line of Figure 1.

2) In addition to the V?V diphones described in our previous report, VV combinations without a glottal stop also occur in Dutch, e.g. in words as 'theater' (/te:ja:tər/). A simple concatenation of /e:j/ and /ja:/ diphones did not result in an acceptable vowel combination, no matter how much the /j/ was speeded up.

3) In Dutch the vowels /e:/ and /o:/ vary in their pronunciation, depending on whether they precede a liquid or another consonant (Cohen et al., 1961). Smoothing a formant value did not yield acceptable vowels. Therefore an additional set of CV diphones was added. These were taken from spoken nonsense words like /bəbo:rə/ and /bəbe:rə/.

4) Concatenation of Vh and hV diphones proved to be impossible because of coarticulation phenomena. VhV triphones were made instead. For the same reason we have incorporated ChV triphones in the inventory.

5) Finally, the inclusion of unstressed /ə/ greatly improved the temporal structure of concatenated utterances.

In its present state, the diphone inventory probably contains some elements which are not really needed. For instance, the V?V triphones can probably all be replaced by V? and ?V diphones. Remember that ?V diphones are already part of the inventory. It is also found from informal listening, that the distinction between consonant clusters occurring within and across morpheme boundaries may be superfluous, except in the case of clusters containing a liquid. If only one CC diphone were retained and used both within and across these boundaries, the number of diphones in the inventory would be greatly reduced.

Improvement in quality

Although the speech obtained with the early system was acceptable, the speech quality could be improved. The aforementioned smoothing algorithm was designed for this purpose. In addition, the quality of each of the separate diphones was checked and polished where necessary. This laborious work could only be carried out by hand since it involved checking for irregularities in formants, bandwidths and amplitude values. Correction of the bandwidth values resulted in a much clearer overall definition, and the vowels and nasals sounded less sharp and more natural.

Further improvement: duration

In the early system, CV and VC diphones were obtained by truncating all the vowels in the middle. This resulted in short vowel parts for diphones containing a voiceless stop consonant, whereas the vowel part of diphones containing a liquid or voiced fricative consonant were relatively long. Concatenation of these diphones sometimes lead to unacceptably short or long vowel duration. The temporal organization of synthetic speech has now been improved by changing the truncation point of the vowel. Instead of taking the middle of the vowel as the truncation point, the diphones have been altered so that the vowel part is identical in duration for all CV diphones for a given vowel. This means that the total duration of a vowel now depends on the VC diphone that follows, or, to be more precise, it depends on the following consonant, as is known to be the case in human speech (cf. Peterson & Lehiste, 1960; Chen, 1970; Nooteboom, 1972; Elsendoorn, 1984).

Work on the temporal structure of synthetic speech at the phonemic and prosodic levels is in progress. A time alignment program written by Niranjan (1984) allows synthetic speech to be automatically aligned with the original utterance. In this way the durational structure of consonant clusters, for instance, can be examined. The regularities that are found can be stored as templates to be used when the diphones are concatenated. In addition, higher-level temporal rules can probably also be derived from these templates.

Temporal structures at the prosodic level can be studied in the same way. It is expected that this method of research will lead to a set of rules that can be included in an automatic speech synthesis system and that these rules will help to improve its quality.

References

Chen, M. (1970) Vowel length variation as a function of the voicing of the consonant environment. Phonetica, 22, 129-159.

Cohen, A., Ebeling, C.L., Fokkema, K. & Holten, A.G.F. van (1961) Fonologie van het Nederlands en het Fries. Den Haag, Nijhoff, 2nd edition.

Elsendoorn, B.A.G. (1984) Tolerances of durational properties in British English vowels. Doctoral thesis, Utrecht University.

Elsendoorn, B.A.G. & Hart, J. 't (1982) Exploring the possibilities of speech synthesis with Dutch diphones. IPO Annual Progress Report, 17, 63-65.

Hart, J. 't & Cohen, A. (1973) Intonation by rule: a perceptual quest. Journal of Phonetics, 1, 309-327.

Niranjan, M. (1984) Time alignment between diphone and natural speech. IPO Internal Report, 471.

Nooteboom, S.G. (1972) The interaction of some intra-syllable and extra-syllable factors acting on syllable nucleus durations. IPO Annual Progress Report, 7, 30-39.

Peterson, G.E. & Lehiste, I. (1960) Duration of syllable nuclei in English. Journal of the Acoustical Society of America, 32, 693-703.