AUTOMATIC DIPHONE SEGMENTATION USING HIDDEN MARKOV MODELS

P. A. Taylor and S. D. isard

Centre For Speech Technology Research,
University of Edinburgh, Scotland.

ABSTRACT - A two stage automatic method of producing a diphone set from nonsense words is described. Firstly hidden Markov models are used to locate phoneme boundaries and then a spectral discontinuity minimisation algorithm is used to choose diphone boundaries.

INTRODUCTION

The speech synthesis system under development at the Centre for Speech Technology Research is based on diphone concatenation. An advantage of diphone synthesis is that different voices may be represented in the synthesizer by simply recording diphone sets from different speakers. In the diphone method of Isard and Miller (1986), a set of nonsense words is recorded from a single speaker, each nonsense word containing at least one diphone. After recording has taken place, the diphones need to be extracted from the nonsense words. Previously this was done by segmenting the nonsense words into phonemes semi-automatically, these segmentations were then checked by hand, and the diphone boundaries were chosen in the appropriate phonemes by a set of simple rules.

Both the phoneme boundary and the diphone boundary procedure may take several man weeks and may prove sub-optimal due to inconsistencies in human segmentation criteria. Other automatic segmentation methods have been proposed by Stella (1985) and van Hemert (1985) and they too involve a stage of hand checking, but it is hoped that our present technique, while failing to be fully automatic, is helpful in that most segmentations do not need to be checked, significantly reducing the number of man-hours spent collecting the diphones. A two stage approach is adopted to tackle the problem. The first (and most problematic) stage is to segment the nonsense words by using hidden Markov models. Using this segmentation information the complete diphone (which contains both phonemes from either half of the diphone in their entirety) is extracted from the nonsense word. No other segmentation is carried out until run time. To choose the precise diphone boundaries, a spectral mismatch minimisation algorithm is used to make the smoothest join between any particular pair of diphones which are to be joined in the utterance under consideration.

HIDDEN MARKOV MODELS FOR AUTOMATIC SPEECH SEGMENTATION

Hidden Markov models (HMMs) can be used to automatically segment speech in much the same way as they can be used for speech recognition. A hidden Markov model system that is set up for segmentation is essentially capable of 'recognizing' a single utterance, with phoneme boundary information being produced as a side effect. A separate HMM is needed for every phoneme (or allophone) and each model is trained on at least ten examples of its phoneme, taken from a set of pre-segmented data from a single speaker. Best results are achieved when trying to segment speech from the same speaker that the models were trained on, and in general the closer the training data is to the data to be segmented (in terms of durational characteristics and phonetic contexts) the better the results will be.

The HMM system used here is similar to that described in McInnes (1990). In this system discrete hidden semi-Markov models are used to include time-durational modelling. This durational modelling adapts HMM transition probabilities so that the overall likelihood of a model being in a particular state conforms to a gaussian distribution with respect to time instead of the exponential distribution inherent in standard HMMs. Eleven cepstral coefficients and three log formant frequencies are used to represent each 20ms frame. A twenty eight dimension vector is used as an initial feature set, where 14 features are used from the preceding frame and 14 features are used from the frame following the current frame. This 28 dimensional vector thus incorporates both static and dynamic information for

the current frame. By use of linear discriminant analysis this vector is reduced to 10 dimensions for vector quantization.

AUTOMATIC PHONEME SEGMENTATION OF NONSENSE WORDS

Diphone Dictionary

The diphone dictionary currently being used contains 2390 diphones from 2169 nonsense words recorded by a single male speaker. Except for some special cases involving schwa and syllabic /l/, /m/ and /n/, all the nonsense words are polysyllabic in form, with the diphone being contained within the stressed syllable. This set of nonsense words was hand segmented into phonemes and from these segmentations the rules of Isard & Miller (1986) were used to choose the best diphone boundary. This operation took several man weeks, and although phoneme boundaries are often unambiguous (and therefore not difficult to segment by hand), diphones boundaries are often arbitrary in that they are a product of 'designed' segmentation criteria, and have no physical characterisation. Inability to enforce physical similarity gives rise to spectral mismatches at diphone joins.

HMM Training

Initially, the HMMs were trained on a set of 200 hand segmented continuous sentences from the same speaker as the diphone nonsense words. Segmentations using these models were found to be unreliable with gross errors often occurring. This was accounted for by the phonetic and durational differences between the nonsense words and the continuous sentences. A set of 5000 hand segmented isolated words also existed for this speaker and from this database a new set of HMMs were trained. This segmentation gave far fewer errors. As 5000 words is rather excessive as a standard training set, a phonetic analysis was made of this set in order to reduce the number of words down to the minimum while still keeping at least 10 examples of each phoneme. The training set was reduced to 400 words this way, and segmentations using this data were found to give as good results as when using the larger set. Fifty six models were used, corresponding to the 44 phonemes of British English, and 12 additional allophones for stops and nasals.

A set of 400 nonsense words were chosen to test the segmentation procedure. These words were automatically segmented and the values for the initial phoneme boundary of the diphone, the final phoneme boundary and the mutual boundary of the two phonemes in the diphone were checked against the hand segmented values.

Segmentation Results

The diphones were grouped into four main classes, consonant-vowel, vowel-consonant, consonant-consonant and vowel-vowel. Each group's segmentations were checked separately to try and identify areas were errors occurred. Results of average error by diphone class and phoneme boundary type are given in table 1.

Van Hemert (1985) claims that an error of 30ms is acceptable for diphone segmentation, and that 96% of the segmentations carried out using his method have boundary errors less than this figure. Although for general speech segmentation purposes 30ms may be considered to be a suitable tolerance factor for phoneme boundaries, for diphone segmentation it is inappropriate to use such rigid criteria. Exact boundary location is not directly important for the initial and final phoneme boundaries as the diphone boundaries will be chosen within these limits.

Diphone type	Initial	Medial	Final
consonant - vowel	16	14	12
vowel - consonant	12	19	17
consonant - consonant	14	16	10
vowel - vowel	12	19	13

Table 1. Average phoneme segmentation errors (ms) per diphone class and boundary type.

As for the mutual boundary between the two phonemes in the diphone, a small error will only have an effect on the *relative* durations of the two halves of the system. This may affect the segmental durations of the system in that it will be impossible to produce a phone that is an exact specified length - the actual length of a phone produced will be the desired length plus or minus the segmentation error. However this effect will only be apparent at the phoneme level, as at syllable and higher levels these effects will cancel out. As seen by table 2 some classes of phoneme proved to be more difficult to segment than others, and most often maximum segmentation errors occurred in the mutual phoneme boundary. In general, segmentation is most difficult where there is most coarticulation, which means that relative phone length will be harder to perceive in the synthetic speech. In other words, errors are most likely to occur where they matter least.

A more important result than the average error is the number of segmentations that have errors over a certain acceptable tolerance, ie those that need to be hand corrected. A study of the nature of segmentation errors helps in reducing the amount of hand checking by identifying areas where the segmentations works reliably and areas where segmentation can be expected to fail. Table 2 shows the number of errors over 30ms and 40ms by phoneme and diphone class. In the test set of 400 nonsense words 95% of the segmentations had errors less than 30ms.

Diphone type	Number of errors > 30ms	Number of errors > 40ms
diphones containing semivowels	6	6
diphones containing nasals	7	6
diphones containing fricatives and stops	4	2
diphones containing only vowels	3	1

Table 2. Number of Errors over 30ms and 40ms by phoneme class, from a test set of 400.

Although phonemes of all classes have errors over 30ms, if the tolerance level is increased to 40ms, nearly all the errors occurred in diphones containing semi-vowels or nasals. In cases where large errors occurred in the mutual boundaries of semi-vowel diphones the coarticulation effect may make

these errors seem less important than at first. However in the case of the larger errors in the nasals, coarticulation is less obvious and these errors may be regarded as more 'genuine'.

DIPHONE BOUNDARY SELECTION

Once the three phoneme boundaries of the diphone have been located the diphone is then stored and the diphones boundaries themselves are only decided at run time. The technique that is used to choose where to place the diphone boundaries is not only automatic, but tries to optimise the boundaries in terms of minimum spectral mismatch between adjacent diphones. Full details of this method are given in Verhoven (1990).

When two diphones are to be joined together the last phoneme of the first diphone will be the same as the first phoneme of the second diphone. The middle thirds of both these phonemes are divided into frames and cepstral coefficients are calculated for each frame. By means of a euclidean distance metric, every frame in the relevant part of the first diphone is compared with every frame in the relevant part of the second diphone, and the positions of the two frames with the lowest score are taken to be where the diphones should be segmented. This technique reduced the amount of spectral mismatch between diphones by 55% when compared to the previous method of using fixed boundaries decided by rules. Informal perceptual results have shown that the synthetic speech produced was judged to be better, especially in dipthongs were human segmentation was very prone to inconsistencies.

The choosing of diphone boundaries based on spectral mismatch minimisation criteria helps to compensate for errors in the initial phoneme segmentation. In most cases there was a high degree of correlation between what was calculated as the middle third in the hand segmented diphones with what was calculated as the middle third in the autosegmented diphones. Unless the diphone boundaries are chosen to be very close to either edge of the middle third of the phoneme, both the hand and the automatic segmented diphones will have the same diphone boundaries. It is therefore difficult to give any exact criteria for what is an unacceptable error in phoneme boundary selection. It is quite possible for the phoneme boundaries to be more than 60ms out and for the minimisation algorithm to still pick the same frame for the diphone boundary. In any case, the diphone boundary selection algorithm can be relied upon to choose the best boundary in the circumstance. If standard diphone boundary selection rules were used, errors in phoneme segmentation would carry over into diphone boundary errors as often these rules are specified in terms of absolute positions within the phoneme.

CONCLUSIONS

The two stage process described here has proved to be an acceptable way of automatically segmenting diphones. The real test of the success of an automatic segmentation method is the amount of hand correction that is required. Nearly all previous 'automatic' segmentations are not truly automatic in that the segmentations must be checked and hand corrections must be made. If this is the case, the amount of time spent creating the diphone set is not much less than when hand segmentation had to be used. Although this technique is not truly automatic, the segmentations failed in easily definable groups and so it is hoped that only these areas will need to be checked and possibly hand corrected. Segmentations involving semi-vowel and nasal diphones would need to be checked as would the occasional cases where diphones occur which contained the same phoneme twice (eg. /m-m/ diphone). This would result in a need to check approximately 15% of the nonsense words.

FURTHER WORK

The automatic segmentation of this diphone set was only possible because the same speaker was used to record the diphone set and the hand segmented isolated word database. However in general it will often be the case that no hand segmented data will be available for a speaker for whom a diphone set is to be made. Research efforts must now be directed at solving this problem. As they stand, hidden Markov models are not particularly good at segmenting speech from speakers who are

different from the training speaker. Automatic segmentation of a speaker's diphone set using models from another speaker would produce significantly worse segmentations than the 95% correct score given above, implying that the entire diphone set would have to be checked. This would defeat the whole purpose of using HMMs as a segmentation method, which was to try and reduce the amount of hand checking to a minimum.

A possible solution would be to auto-segment a small set of words from a new speaker, then re-train the models using these segmentations and use these new models to segment the diphone set. Thus only the training set would have to be checked. It should also be possible to improve on the training data by selecting data closer to the words that are to be segmented. A large improvement was gained by using isolated words instead of continuous sentences for the training data and it may be possible to improve results by using words of even greater similarity to the nonsense words. A possible option would be to train on selected nonsense words themselves and therefore capture the characteristics of isolated *nonsense* words.

The cepstral vectors used for the diphone boundary selection performed well and undoubtedly improvements in speech quality were obtained. However it may be worthwhile investigating what other features may be used to compare frames of speech for optimised joining. Both formant values and mel scaled data are currently be assessed for their suitability.

ACKNOWLEDGMENTS

We would like to thank Fergus McInnes, Steve Hiller and Gordon Watson for their help with the autosegmentation experiments.

REFERENCES

Isard, S.D., Miller (1986) *Diphone Synthesis Techniques*, IEE Conference Publication no 258, p77-82.

McInnes, F.R., McKelvie, D., Hiller, S.M. (1990) *The Structure, Strategy and Performance of a Modular Continuous Speech Recognition System,* Proceedings, IOA Autumn Conference.

Stella, M. (1985) *Speech Synthesis,* in Fallside & Woods (eds), Computer Speech Processing, London: Prentice Hall, p 421 - 460.

van Hemert, J.D. (1985) *Automatic Diphone Preparations*, IPO annual Progress Report 20, p 23 - 32.

Verhoven, J.W.M (1990) *Context Sensitive Diphones as Units in Speech Synthesis,*Proceedings, IOA Autumn Conference.