



# A hidden Markov-model-based trainable speech synthesizer

R. E. Donovan<sup>†‡</sup> and P. C. Woodland<sup>§</sup>

*Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ U.K.*

---

## Abstract

This paper presents a new approach to speech synthesis in which a set of cross-word decision-tree state-clustered context-dependent hidden Markov models are used to define a set of subphone units to be used in a concatenation synthesizer. The models, trees, waveform segments and other parameters representing each clustered state are obtained completely automatically through training on a 1 hour single-speaker continuous-speech database. During synthesis the required utterance, specified as a string of words of known phonetic pronunciation, is generated as a sequence of these clustered states using a TD-PSOLA waveform concatenation synthesizer. The system produces speech which, though in a monotone, is both natural sounding and highly intelligible. A Modified Rhyme Test conducted to measure segmental intelligibility yielded a 5.0% error rate. The speech produced by the system mimics the voice of the speaker used to record the training database. The system can be retrained on a new voice in less than 48 hours, and has been successfully trained on four voices.

© 1999 Academic Press

---

## 1. Introduction

In recent years, the application of statistical modelling techniques to the problem of automatic speech recognition has proved to be very successful and most speech recognition systems are now based on these techniques to some degree. In particular, hidden Markov models (HMMs) have been widely used to model the acoustics of the speech signal, to relate vectors of spectral parameters to the linguistic concepts of phonemes and words. These techniques have proven so successful because they enable detailed self-consistent models to be constructed using large amounts of data; much more data than might be analysed by hand to construct a rule-based system for example. The models are also often relatively easy to optimize, in some useful sense, and can be rapidly retrained on new databases to model new speakers or new languages.

Most current speech synthesis systems, however, whether using rule-based formant synthesis (Klatt, 1982; Allen, Hunnicutt & Klatt, 1987; Hallahan, 1996), or more recently concatenation synthesis (Olive, 1990; Sorin, 1994), generally use rules and/or concatenation units generated by hand. The research described in this paper was conducted to determine whether

<sup>†</sup>R. E. Donovan is now at the IBM T. J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, U.S.A.

<sup>‡</sup>E-mail: [red@watson.ibm.com](mailto:red@watson.ibm.com)

<sup>§</sup>E-mail: [pcw@eng.cam.ac.uk](mailto:pcw@eng.cam.ac.uk)

statistical techniques similar to those applied so successfully to the problem of speech recognition could be similarly applied to the problem of speech synthesis. Specifically, the aim of this research was to build a concatenation synthesizer which could obtain all the parameters required to synthesize speech through training on a suitable database. The research focused on the generation of speech in a monotone given a word string of known pronunciation; text-to-phoneme conversion and fundamental frequency contour generation were not attempted.

The advantage of constructing a concatenation synthesizer using statistical data-based techniques is that it should be possible to select a superior set of concatenation units than are used in current systems, and hence generate better quality synthetic speech. In addition, such a system could be rapidly retrained on new data to synthesize new voices or new languages.

This paper is structured as follows. Section 2 presents a review of concatenation synthesis, discussing both traditional hand-generated unit inventories and recent research into automatic methods of selecting and segmenting units. Section 3 gives a brief overview of the system described in this paper. Section 4 describes the hidden Markov model system used in the current research in the first stage of unit inventory generation. Section 5 describes the second stage of this process, and the estimation of other parameters necessary for synthesis, which is itself described in Section 6. Section 7 presents a description of the speech produced by the system, the results of listening tests conducted on the speech and a discussion of why the system performs as it does and of the problems remaining. Finally, possible future work is discussed in Section 8 and the conclusions of this paper are presented in Section 9.

## 2. Concatenation synthesis inventories

For text-to-speech synthesis, with a potentially unlimited vocabulary, the task of pre-recording all possible words for use in a word concatenation system is prohibitive and therefore some form of subword concatenation unit must be used from which all words can be constructed as required. Traditionally, concatenative speech synthesis systems used a set of synthesis units all of the same type. In English and other European languages, the diphone (or dyad) was often the unit of choice (Dixon & Maxey, 1968; Courbon & Emerard, 1982), although demi-syllables have also been used. Diphones are segments of speech which begin in the middle of one phone and end in the middle of the next, and the unit endpoints are therefore often in regions of relative spectral stability. These regions are often spectrally similar for a given phone across phonetic contexts, and therefore diphones often concatenate relatively smoothly. The diphones used in synthesis systems were typically segmented by hand from nonsense words specially prepared to contain the required unit, usually in a neutral phonetic context. Preparation of new databases was therefore slow and time consuming.

As digital storage problems eased, and concatenative systems were more thoroughly researched, augmented diphone systems were introduced (Olive, 1990; Bigorgne *et al.*, 1993; Lernout & Hauspie, 1999). In these systems, longer polyphone units were introduced to improve concatenation smoothness in those contexts in which diphones joined least smoothly. The resulting systems performed much better than their more simple predecessors. However, the selection of which longer units to use was performed with considerable human intervention, and was therefore probably not well optimized.

Recently, research has been conducted into methods of automatically selecting and segmenting concatenation synthesis unit inventories. The remainder of this section will briefly review this research.

All methods of selecting or segmenting synthesis unit inventories require a knowledge of the phone sequence of the training database. This transcription was traditionally generated

by human experts. However this is not ideal, since human transcription is both slow and subject to disagreement between transcribers (Ljolje & Riley, 1993). At the time of writing, unconstrained phone recognition does not provide a realistic solution to the transcription problem. Other automatic solutions, such as concatenating pronunciations obtained from a pronunciation dictionary (Brugnara, Falavigna & Omologo, 1992) or using the text-to-phoneme conversion algorithms of a text-to-speech system, provide a solution, but since they do not reference the acoustic data of the database are not likely to be very accurate. Currently, the most accurate automatic method of obtaining phonetic transcriptions, and the method used in the current work, is to use a hidden Markov model system to select between alternative pronunciations on the basis of the acoustic data. This is the method usually used to obtain the correct model sequence for parameter re-estimation during the training phase of HMM-based speech recognition systems (Woodland *et al.*, 1994).

Automatic phone and diphone segmentation have been increasingly investigated in the last several years. Phone segmentation was initially only useful in synthesis as a precursor to diphone segmentation. However, the increases in speed and self-consistency which it offers also make it attractive for use with the phone-sized unit selection algorithms discussed below. Most recent phone segmentation systems use hidden Markov models (Ljolje & Riley, 1991; Taylor & Isard, 1991; Brugnara *et al.*, 1992; Ljolje & Riley, 1993; Boëffard, Cherbonnel, Emerard & White, 1993). The differences between these systems were whether they used context-dependent or context-independent models, manually or automatically determined phone boundaries during training, continuous speech or isolated words, and single-speaker or multi-speaker databases. The quoted results are somewhat difficult to compare, but it appears that a good system could place 80% of boundaries within 11.5 ms of manually placed boundaries when segmenting single-speaker databases (Ljolje & Riley, 1993). This result is encouraging, since it compares well with the result of 80% of boundaries being within 8 ms of each other when comparing two human segmentations of the same speech (Ljolje & Riley, 1993). Recent attempts at automatic diphone segmentation (Taylor & Isard, 1991; Boëffard, Miclet & White, 1992) have used a hidden Markov model system to perform an initial phone segmentation before attempting diphone segmentation. The latter was performed either in the traditional manner by placing boundaries in regions with low spectral derivatives, or by determining variable boundaries for each diphone so as to minimize the spectral concatenation discontinuity for each diphone pair. The variable boundaries could either be pre-computed or determined dynamically during synthesis. Synthetic speech produced using variable boundaries was reported to be no worse than that produced using fixed boundaries (Boëffard *et al.*, 1992), or even somewhat better (Taylor & Isard, 1991).

As discussed above, diphone units, and specific polyphone units, are often used in concatenation synthesis systems because they concatenate relatively smoothly. However, the use of these units represents a human imposition of the level of context information which can affect the sequence of segments concatenated to synthesize a new utterance. This situation is non-ideal, since the realization of phonemes can be affected systematically by contexts not captured in diphones, and for which it is difficult to manually select an optimal set of polyphone units. To address this problem research has been conducted into automatic unit selection algorithms. This research has generally been performed either with phone length units or with subphone units.

The automatic selection of phone length units has pursued two main themes: systems which use decision tree clustering and those which do not. The decision tree clustering approach to unit selection for speech synthesis was pioneered Nakajima and Hamada (1988) for Japanese, and extended to English by Nakajima (1993) and Itoh, Nakajima and Hirokawa (1994). It was

also investigated by Wang, Campbell, Iwahashi and Sagisaka (1993), and forms the basis of the work described in this paper. In this approach, all the instances of a given phoneme in a single-speaker continuous-speech database are clustered into equivalence classes according to their phonetic and wider contexts. The decision trees which perform the clustering are constructed automatically so as to maximize the acoustic similarity within the equivalence classes. This approach is similar to that used in modern speech recognition systems to generate hidden Markov models in different phonetic contexts (Lee *et al.*, 1990; Bahl, de Souza, Gopalakrishnan, Nahamoo & Picheny, 1991). In the synthesis systems, parameters or segments were then extracted from the database to represent each leaf in the tree. In synthesis the trees were used to obtain the unit sequence required to produce the desired sentence. A key feature of this method is that the tree construction automatically determines which context effects are most important in terms of their effect upon the acoustic properties of the speech, and thus enables the automatic identification of a leaf containing segments or parameters most suitable for synthesizing a given context during synthesis. Furthermore, this is true even when the context required was not seen in training.

Systems which do not use decision trees have been investigated by Iwahashi, Kaiki and Sagisaka (1992), Hauptmann (1993), Black and Campbell (1995) and Hunt and Black (1996). In the ATR  $\nu$ -Talk system (Iwahashi *et al.*, 1992) Japanese speech was synthesized by concatenating variable length segments, each being an integer number of phones in duration, selected from a continuous speech database. The synthesis time selection proceeded by first determining the size of the units to be concatenated by using a dynamic programming search to simultaneously minimize the expected perceptual concatenation degradation between successive phones and the Contextual Spectral Difference (CSD). The CSD is a measure of the spectral difference between speech segments in the database in the triphone context to be synthesized (the target context) and segments in the database in other triphone contexts which may be used instead during synthesis. Methods were developed for estimating the CSD between target and alternative contexts even when the target context did not exist in the training data. Once the size of the units to be concatenated had been determined another dynamic programming search was conducted to select the segments to actually use to produce each unit during synthesis. This search minimized the spectral discontinuity between successive units allowing for optimal boundary placement in the process. Fundamental frequency, duration and energy were predicted using separate modules and the synthetic waveform generated using a Log Magnitude Approximation filter.

The SpeakeZ system described by Hauptmann (1993) was somewhat simpler, and was used to synthesize speech in English. Unlike the  $\nu$ -Talk system, and the CHATR system described below, fundamental frequency, duration and energy were not supplied by any external means. Instead, the natural prosody of the selected segments was simply retained. In the SpeakeZ system, the sequence of phone length segments used to construct each synthetic sentence was selected at a synthesis time from a very large single speaker speech database. The selection was performed using a manually defined heuristic which calculated a context-matching score between each phone required in synthesis and all occurrences of that phone available in the database. The heuristic included different levels of context information, such as stress, phonetic context, and word and utterance boundary information, which were manually ranked in order of their importance. The segments with the best scores were then concatenated using the PSOLA (Pitch Synchronous Over-Lap and Add) algorithm (Charpentier & Stella, 1986; Moulines & Charpentier, 1990), to smooth the fundamental frequency only at concatenation boundaries.

In Black and Campbell (1995) the authors, also at ATR, stated that the  $\nu$ -Talk system did

not perform well when applied to synthesis in English due to the larger number of phonemes and more variable prosody than are found in Japanese. To overcome this problem a more general algorithm was therefore investigated (Black & Campbell, 1995; Hunt & Black, 1996) as part of the CHATR system. In this system the target specification for a synthetic utterance consisted of the phoneme sequence to be synthesized, and each phone's desired fundamental frequency, duration and energy. The sequence of phone length segments used in synthesis was selected from a large single speaker speech database using a dynamic programming algorithm to minimize a cost function which included both context-matching criteria and concatenation continuity considerations. Unlike the  $\nu$ -Talk system, the context-matching criteria, measured between the target context and the contexts available in the database, included both phonetic and prosodic context-matching costs. The use of variable length segments was encouraged by specifying that the continuity cost between segments adjacent in the database was zero. Methods were developed to enable the weights of the different context factors within the cost function to be determined automatically, rather than by hand.

Numerous investigations have also been conducted into the use of subphone units for concatenative speech synthesis. The use of these units in speech synthesis has arisen largely because they are used in finite state speech modelling methods, such as vector quantization (VQ) and hidden Markov models (HMMs). These systems model speech by quantizing it into a finite number of acoustically self-similar states. Most work in the literature has used either VQ techniques or ergodic HMM systems.<sup>1</sup> These systems are directly applicable to vocoding, or speech compression, since only the state sequence of the original speech must be transmitted or stored. However, using such systems for text-to-speech synthesis is less straightforward since the relationship between words or phonemes to states must be established and this is non-trivial. Nevertheless, some success has been reported in the literature with systems in which a discrete HMM was used to relate phonemes to the acoustic states (Giustiniani & Pierucci, 1991; Sharman, 1994). More recently, work has been reported which used multiple model HMM systems in which the relationship between phonemes and states was explicit (Tokuda, Kobayashi & Imai, 1995a; Tokuda, Masuko, Yamada, Kobayashi & Imai, 1995b). In this system, synthesis was achieved by concatenating phone HMMs in the order defined by the sentence to be synthesized to create a composite model. The most likely sequence of observations (cepstral vectors) to be generated by this model was then determined using an iterative algorithm. The inclusion of delta coefficients in the feature vectors led to a smoothly evolving sequence of generated vectors from which the speech was then synthesized directly.

The automatic unit selection algorithms described above have all worked with phone length or subphone length segments of speech. These units have the disadvantage that the boundaries between them are often in regions of rapid spectral change making smooth concatenation more difficult to achieve than with, for example, diphone units. However, the hope is that careful unit selection coupled with self-consistent automatic unit segmentation will overcome this disadvantage and that the use of a large number of automatically selected units will lead to synthetic speech superior to that obtained using manually selected diphones and polyphones.

### 3. System overview

The speech synthesis system described in this paper uses automatic statistical methods to select and segment a set of synthesis units from a large speech database. The entire training

<sup>1</sup>Ergodic HMMs are those in which any state can be reached from every other state in a finite number of transitions.

process is fully automatic and can be completed in less than 48 hours from the commencement of recording of the database. Speech synthesis is achieved by concatenating appropriate units in the order necessary to produce the desired synthetic sentence.

The speech database contains approximately 1 hour of speech recorded from a single speaker in the form of continuous read speech sentences with normal prosody. The speech is recorded together with a laryngograph signal which is used to determine the moments of glottal closure through voiced regions of the speech. Knowledge of these is required by the signal-processing algorithm, TD-PSOLA (Time Domain Pitch Synchronous Over-Lap and Add), which is used to concatenate segments during synthesis.

A set of speaker-dependent hidden Markov models (HMMs) is trained on the database and used to segment it into HMM-state-sized units. The HMMs used are cross-word decision-tree state-clustered context-dependent models. For each state position in the (mostly) three-state left-to-right HMM modelling each of the (approximately) 50 phones, a decision tree is constructed to cluster the training speech into acoustically self-similar clustered states by splitting the data according to its phonetic context labels. The tree-growing process automatically determines which context effects are most important at each stage in the tree's construction. The (approximately) 150 trees thus provide a mechanism for reaching the most appropriate clustered state given a new context in synthesis, even for contexts not seen in training. In the synthesis system, the decision-tree clustered HMMs are used only to segment the training data into a sequence of clustered states. The decision trees are also used to define the system's synthesis units. A segment selection algorithm is used to select a single segment from all the waveform segments in a given clustered state to represent that state in synthesis. In addition, duration and energy parameters are estimated from the training data for each clustered state in the system.

In synthesis the required sentence, specified as a string of words of known phonetic pronunciation, is first converted to a phone sequence, and then, using the decision trees, to a clustered state sequence, and hence a segment sequence. The segments are then concatenated using a TD-PSOLA synthesizer to scale each segment to the energy and duration values estimated for its clustered state during training, and to modify each segment's fundamental frequency to produce the required  $F_0$  contour.

## 4. The HMM system

### 4.1. Training data

The system was trained on approximately 1 hour of continuous speech recorded from a single speaker, read from a popular novel. Ideally, the training text should reflect the task for which the synthesizer is to be used, since it is likely that using training text similar in style and content to that to be synthesized ensures an appropriate balance of acoustic information across contexts likely to be encountered during synthesis. Those contexts which occur most frequently in training are modelled most accurately, which is appropriate since they are likely to be those required most frequently in synthesis. Less frequently seen contexts are modelled less accurately, but this is acceptable since they are less frequently used in synthesis. However, problems can arise, and have been found in later work, when contexts are required in synthesis which are very different from anything seen in training. For this reason some degree of database preparation to ensure a minimum coverage of possible contexts is probably wise, though it was not performed in the work described in this paper.

Four databases were used to train the system described in this paper, which will be termed

the M2, M3, F1 and F2 databases.<sup>2</sup> The M labels refer to male databases and the F labels to female databases. All the speakers used spoke British English. All databases except the F1 database included laryngograph signals (see Section 5.2). The system was developed principally using the M2 database, and all database specific quantities mentioned in this paper refer to this database.

#### 4.2. Context-independent models

The initial acoustic analysis parametrized the training data into mel frequency cepstral coefficients, energy, and their first and second differentials using a 25 ms frame duration with a 6 ms frame shift between frames (hereafter referred to as a 25/6 analysis). A set of three-emitting-state left-to-right no-skip context-independent hidden Markov models was trained using embedded Baum–Welch re-estimation using the orthographic transcriptions of the training data and a pronunciation dictionary to determine the phonetic transcription of each training sentence. Where multiple pronunciations existed one was initially selected at random, but was later replaced by the most likely pronunciation determined using the partially trained models.

Once trained, the context-independent models were altered in several ways. Firstly, the plosives /b/, /d/, /g/, /p/, /k/, /t/ and the affricates /ch/ and /jh/ were each split into a one-state closure phone followed by an optional one-state burst phone. This was done in order to enable the accurate modelling of both released and unreleased plosives which had been seen to occur in the same immediate phonetic context. This method was preferred over the alternative method of introducing skip transitions into the plosive transition matrices, since it enabled a neighbouring burst to be distinguished from a neighbouring closure during clustering. Silence was modelled using a seven-state left-to-right model in order to enforce a longer minimum silence duration when using the new data analysis described below. This was necessary to prevent the erroneous insertion of extra silences by the system.

The data was reanalysed using frame sizes and shifts determined by the class of phone aligned in a first pass. Regions of voiced speech were analysed at the same size and shift as before in order that the frames averaged over several glottal pulses to reduce frame-placement effects on the parameters obtained. For unvoiced speech and silence these considerations did not apply and therefore a shorter frame size of 6 ms was used, with a 4 ms frame shift, or a 2 ms shift in the case of plosives. The reanalysis was introduced principally to improve the modelling of plosives, in which the 25 ms frame size combined with the short timescales and rapid transitions involved often meant that as many frames lay partially within the closure or burst regions as lay wholly within them. Reanalysing with shorter frames helped to ensure that a larger proportion of frames lay wholly within each region which could therefore combine to produce “purer” models and hence produce improved frame-state alignments.

The lexical stress markings in the dictionary were used, where sufficient data existed,<sup>3</sup> to estimate models of vowels with primary, secondary, and no stress. Where there was insufficient data either the primary and secondary stress levels remained tied, or all three levels remained tied. The distinct models were estimated at this stage to aid the selection of correct pronunciations from the dictionary. During clustering, stress labels were used as additional clustering features, as described in Section 4.3.

In a similar way, where sufficient data existed, models were created representing the syllabic and non-syllabic instances of /n/, /m/, and /l/. In the case of /l/, an additional model representing

<sup>2</sup>The male labelling begins at M2 in order to be consistent with the labels used in Donovan (1996a).

<sup>3</sup>Deemed, somewhat arbitrarily, to be more than 50 occurrences of each stress level.

the dark /l/ was also created. The syllabic labels were also used as additional clustering features during clustering (see Section 4.3).

Several other alterations were also performed to the models and alignments. These included the replacement of many closures by a dummy closure model which was then ignored when determining contexts during clustering and not used to estimate any synthesis parameters. The closures replaced were those present in the transcriptions which had suspiciously short alignments, deemed to be less than 6 ms, which were often not actually present in the speech of the database. The dummy model was used in preference to simply removing the closures from the transcriptions to avoid problems in synthesis caused by alignment errors with neighbouring phones when closures were erroneously removed. Also, a heuristic was introduced to remove short silences, which were always optionally present between words, when they were aligned to the closures of words beginning or ending with a closure. For further information on other minor adjustments see Donovan (1996a).

The new models were retrained using the parameters from the new acoustic analysis, new alignments were then generated, and the whole reanalysis procedure repeated in order that any changes in the alignments could affect the analysis finally used. Finally, after further retraining, position in word labelling (word beginning, word ending or word medial position) was also introduced. All tied context-independent models were then untied and used to initialize a context-dependent model, dependent on the central and immediately adjacent phones, for every distinct context in the training data. At this stage there were 286 context-independent models, leading to a total of over 23 million possible context-dependent models, 13 thousand of which were present in the training data. The context-dependent models were then retrained prior to clustering.

#### 4.3. Context clustering

Corresponding states from each of the 52 base phones were clustered using binary decision trees. A base phone is defined to be one of the phone set used in the pronunciation dictionary with the plosives split into distinct closure and burst phones. An early (unnamed) version of the British English Example Pronunciations Dictionary was used. For each base phone, all corresponding states were pooled into the root node and, together with occupation counts saved from a preceding re-estimation, used to calculate the likelihood of the data fitting a single Gaussian in that node. A set of context questions was then used to suggest splits of the data into two child nodes. The likelihood of the data fitting a single Gaussian in each of the nodes was calculated for each possible split, and the question which resulted in the biggest gain was used to perform the actual split (Young, Odell & Woodland, 1994).

*Author: please define in full*

The questions used are a modified version of the set developed for use in the HTK large vocabulary speech recognition system (Woodland, Odell, Valtchev & Young, 1994) which are listed in Odell (1995). The original questions grouped phonetic contexts using broad context questions, such as “Fricative?”, “Front Vowel?”, “Nasal?”, etc. They were modified for this work to reflect the phone set being used, and extended to include questions about stress level, the syllabic nature of some phones, and position in word information. These multi-dimensional context questions were included simply by duplicating the original questions with the new context dimension (e.g. stress level) both as a factor and not as a factor. The result was a very large number (1928) of questions. In this work, the questions were applied only to the immediate (triphone) phonetic context, with questions about stress level, syllabic nature, and position in word also being asked about the central phone.

The new clustering features and their associated questions were added for several reasons.

Syllabic labelling and clustering was introduced to provide a mechanism by which segments representing different syllable-dependent allophones in the same immediate phonetic context could be distinguished, and so clustered into different clustered states if significantly different. Without this distinction such segments were always clustered into the same clustered state and since only one segment was retained to represent each clustered state during synthesis, the same segment would be used to produce every allophone. Worse still, since the segments associated with each state of a three-state model were selected independently, in synthesis consecutive segments could come from different allophones. The resulting formant discontinuities, for example from the concatenation of a dark /l/ segment, a light /l/ segment, and a syllabic /l/ segment, were clearly audible. Position in word clustering was introduced because there was otherwise no mechanism by which a word internal context could be distinguished from a cross-word context, and because the research of Nakajima (1993) had shown that word boundary information could be useful during clustering. Stress level clustering was also shown to be important by Nakajima (1993) and by Wang *et al.* (1993). The usefulness of the new clustering features was confirmed in the current work. It was found that questions referring to both stress level and contexts with specific within-word positions were used to split the root node during the construction of some trees. In addition, a question about the syllabic nature of /l/ was asked as the first question of the tree built for the leftmost /l/ state.

The stopping criterion used to limit tree growth was to insist on a minimum number of 12 occurrences of each clustered state in the training data. This differed from the approach used in HTK recognition systems where a minimum number of frames of data is usually used. The change was made due to the small amount of training data used. Using a minimum number of occurrences enforces a useful minimum number of frames, while the reverse is not true. The alternative, using a minimum number of frames criterion was tried, but sometimes resulted in only one or two occurrences of a clustered state in the training data, which was undesirable for several reasons. Firstly, such a state could begin to model speech not matching its linguistic label, since the model was not constrained to match other occurrences, and secondly, too few occurrences were available to the subsequent segment selection algorithms. The number 12 was chosen to enforce a minimum number of frames per clustered state significantly larger than the five frames per clustered state shown to give poor results in Donovan and Woodland (1995), and yet to yield several thousand clustered states in order to have more units in the system than in a simple diphone system. The exact value was not thoroughly investigated; a value of 12 led to approximately 5700 clustered states.

The state-clustered context-dependent models were further retrained and then used to obtain a final state alignment of the training database.

## 5. Synthesis parameter estimation

The next stage of system construction was to determine parameters for each clustered state to enable that state to be synthesized. The time domain (TD) version of the PSOLA synthesis technique (Charpentier & Stella, 1986; Moulines & Charpentier, 1990) was selected for use in synthesis, since it offered the possibility of very high quality synthesis and was relatively easy to implement. Its use required only that one segment be selected to represent each clustered state during synthesis and that the moments of principal excitation of the vocal tract be determined in the voiced sections of the segments to be used. In addition, average duration and energy parameters were estimated from the training data for each state.

### 5.1. Segment selection

The clustering procedure described in Section 4.3 resulted in at least 12 segments of speech being associated with each clustered state in the system. In principle, all the segments associated with each clustered state could be available for concatenation at synthesis time, with the particular segment to use being chosen dynamically during synthesis. This possibility is discussed further in Section 8. However, in this work, one particular segment was selected during system construction to represent each state. The segment selection algorithm consisted of three stages:

**Stage (i).** Discard all the occurrences of a state with a duration shorter than 80% of the average duration of that state.

This was beneficial for two reasons. Firstly, it made the selection of segments which were present in the transcriptions but which were absent from the speech unlikely, since such segments tended to have short alignments. Secondly, it ensured that when durations were stretched during synthesis (see Section 5.3), stretching factors averaged less than 1.52 even when synthesizing isolated words. This factor was approximately the limit of the TD-PSOLA implementation's ability to increase durations without distortion.

**Stage (ii).** Discard all the occurrences of a state with an average short-term energy per sample (*s.t.e.p.s.*)<sup>4</sup> lower than 80% of the average *s.t.e.p.s.* of the speech from all the occurrences still under consideration.

During synthesis each segment was scaled to have a *s.t.e.p.s.* equal to the average for the clustered state that it represented. Ensuring that relatively high energy segments were selected was beneficial since it ensured that segments were not scaled up by very large factors during synthesis. Such scaling could otherwise introduce artifacts into the synthetic speech when non-speech sounds or speech irregularities were produced at energies far higher than those at which they occurred.

**Stage (iii).** Select the occurrence still under consideration which has the highest average log-likelihood per frame in the state alignments.

This ensured that the segment selected was that most likely to be observed using the state Gaussian, and was therefore likely to be a representative segment for that state, rather than an outlier.

### 5.2. Pitch-mark identification

The determination of the moments of principal excitation of the vocal tract (usually the moments of glottal closure), was performed using a commercially available program called *Epochs* from the Entropic Research laboratory. *Epochs* uses a dynamic-programming algorithm and a set of costs and rewards to determine the most likely set of moments in its input signal. Initially, and throughout for the F1 database, the linear prediction residual of the speech was used as the input signal. Later however, superior results were obtained by using a processed form of a laryngograph signal (Fourcin & Abberton, 1971) recorded along with the speech.

A laryngograph is a device which measures the impedance between two electrodes held to the neck either side of the larynx. The signal recorded reflects the status of the vocal cords, since when the cords are open current cannot flow across the air gap. The differential of the signal therefore reflects the rate of change of the status of the vocal cords, and usually has large spikes at the moments of glottal closure. It was this signal, thresholded at 1.6–2.0 times

<sup>4</sup>The *s.t.e.p.s.* of the sequence of waveform samples  $y_1$  to  $y_N$  is  $\frac{1}{N} \sum_{n=1}^N y_n^2$ .

the root-mean-square (r.m.s.) value of the signal-to-remove noise, which was used as the input to *Epochs*. The number of errors in determining the moments of closure was typically a few per sentence for male speakers, but rather more for some female speakers.

The moments of glottal closure, as determined above, become the voiced pitch-marks (V-marks) used by the TD-PSOLA algorithm. Unvoiced marks (U-marks) were added at a uniform spacing (4 ms) to enable duration modifications to be performed on unvoiced speech. The V-mark times had to be adjusted to allow for the time lag between the laryngograph signal measured at the vocal cords and the speech signal measured a few centimetres from the lips. The lag is determined by the distance between the two points of measurement (and the speed of sound), and is therefore a little different for males and females. It was measured to be between 500  $\mu$ s and 690  $\mu$ s for the speech of one male speaker and one female speaker. Since the corrections required were small and fairly similar, a single correction of 562.5  $\mu$ s (nine samples at 16 kHz) was used in all cases.

### 5.3. Duration parameters

During synthesis each state  $s$  was synthesized for a duration  $\tau_s$  given by

$$\tau_s = \mu_s + f * \sigma_s, \quad (1)$$

where  $\mu_s$  is the mean duration of state  $s$  in the training data,  $\sigma_s$  is the standard deviation of the duration of state  $s$  in the training data, and  $f$  is a state-independent scaling factor.

The scaling factor  $f$  was set to either 0.1 in the case of continuous speech, or 0.5 in the case of isolated words. This slowing of the speech was necessary because the durations extracted from the database were those of fluent continuous natural speech. Continuous synthetic speech produced using these durations was sometimes too fast to understand, and was much too fast when synthesizing isolated words. These scaling factors corresponded to synthesizing each state for an average of 1.04 times and 1.22 times its average duration, respectively. The 80% duration threshold used in the segment selection algorithm described in Section 5.1 therefore meant that stretching factors averaged less than 1.30 and 1.52 in the two cases. These stretching factors were approximately within the acceptable range for the TD-PSOLA implementation used.

Equation (1) ensured that when the speaking rate was altered during synthesis, those states which were seen to vary most in duration in the database were varied the most, and those states which were seen to have fairly constant durations in the database were varied the least. This approach would be justified if all the duration variation seen for a particular state in the database was due to local variations in speaking rate since the deviation figure would then simply reflect this variation. In practice, the deviation figure undoubtedly also reflected the different durations associated with the different contexts clustered into each state, in which case equation (1) was perhaps less appropriate. However, despite this caveat, the method seemed to give reasonable results.

### 5.4. Energy parameters

All the speech labelled as belonging to a particular clustered state was pooled to calculate the mean short-term energy per sample (s.t.e.p.s.) for that state. During synthesis each segment was scaled to the mean s.t.e.p.s. for its state. This procedure provided some degree of energy smoothing during synthesis, since the mean s.t.e.p.s. figures of consecutive states were likely to be more similar than the s.t.e.p.s. figures of segments selected to represent those states.

## 6. Synthesis

During synthesis the words of the utterance to be synthesized were first converted to a phone string by dictionary look-up. Where multiple pronunciations existed one was chosen manually. The only text processing included was to interpret question marks, exclamation marks, full stops, and commas as short durations of silence. The phone string was then converted to a sequence of context-dependent phones, and this to a sequence of clustered states using the decision trees. The decision trees enabled clustered states to be assigned to all possible contexts, whether or not they were seen in the training data. The waveform segments associated with the sequence of clustered states were then concatenated using an implementation of the TD-PSOLA algorithm. The duration of each segment was scaled to the value computed using equation (1), and the energy to the mean s.t.e.p.s. figure of the appropriate state. In general, the synthetic  $F_0$  contour was constant, set to the average  $F_0$  of the speaker used to record the training data. However, some experiments were conducted using stylized and transplanted  $F_0$  contours.

The implementation of the TD-PSOLA algorithm is described in detail in Donovan (1996a). The implementation was shown to work very well for  $F_0$  changes of factors of 1.2 or less, and for duration compression up to a factor of at least 2.0. It also worked reasonably well when raising  $F_0$  up to a factor of 2.0, up to a factor of about 1.5 when lowering  $F_0$ , and up to a factor of about 1.5 when lengthening durations. Beyond these limits, artifacts began to appear in the synthetic speech. In the case of raising  $F_0$  or lengthening duration these artifacts were due to the over-repetition of unvoiced/mixed-excitation waveform segments, which introduced an artificial periodicity into the synthetic speech. In the case of lowering  $F_0$  they were due to the size of the Hanning windows used, which were fixed at twice the synthetic fundamental period, containing strong images of the original  $F_0$  causing a mis-match between this and the synthetic  $F_0$  in the synthetic speech.

## 7. Results

The final system was trained on four databases. Each database took about 4 hours to record, after which it took about 40 hours of computer time (as a single user on an HP735-99) to construct a synthesis system. Once trained the system could synthesize speech in a monotone from a word string specification of known pronunciation. The synthetic speech mimicked the voice of the speaker used to record the training database.<sup>5</sup>

### 7.1. Synthetic speech

Figure 1(a) and (b) shows the speech waveform and wideband spectrogram of the sentence fragment “When a sailor in a . . .”, taken from a synthesized version of the part-sentence “When a sailor in a small craft faces the might of the vast Atlantic Ocean today, . . .”. The speech was synthesized using a system trained on the M2 database. Figure 1(c) was obtained from a natural version of the same utterance spoken by the speaker used in the M2 database. There are more glottal pulses in the synthetic speech spectrogram because the synthetic speech is slower than the natural speech. In the first half of Figure 1(b) the synthetic speech spectrogram can be seen to be quite similar to that of the equivalent natural speech. Formant discontinuities at state boundaries are sufficiently small that the state boundaries are often difficult to identify visually in the spectrogram or audibly in the speech. In the second half of Figure 1(b), however, the

<sup>5</sup>Audio examples of the synthetic speech produced by the system are available from Donovan (1996b).

TABLE I. The Modified Rhyme Test error rates obtained by Logan *et al.* (1989)

System	Error rate (%)
Natural Speech	0.53
DECtalk 1.8, Paul	3.25
DECtalk 1.8, Betty	5.72
Prose 3.0	5.72
MITalk-79	7.00
Amiga	12.25
Infovox SA 101	12.50
TSI-Proto 1	17.75
Smoothtalker	27.22
Votrax Type 'n' Talk	27.44
Echo	35.56

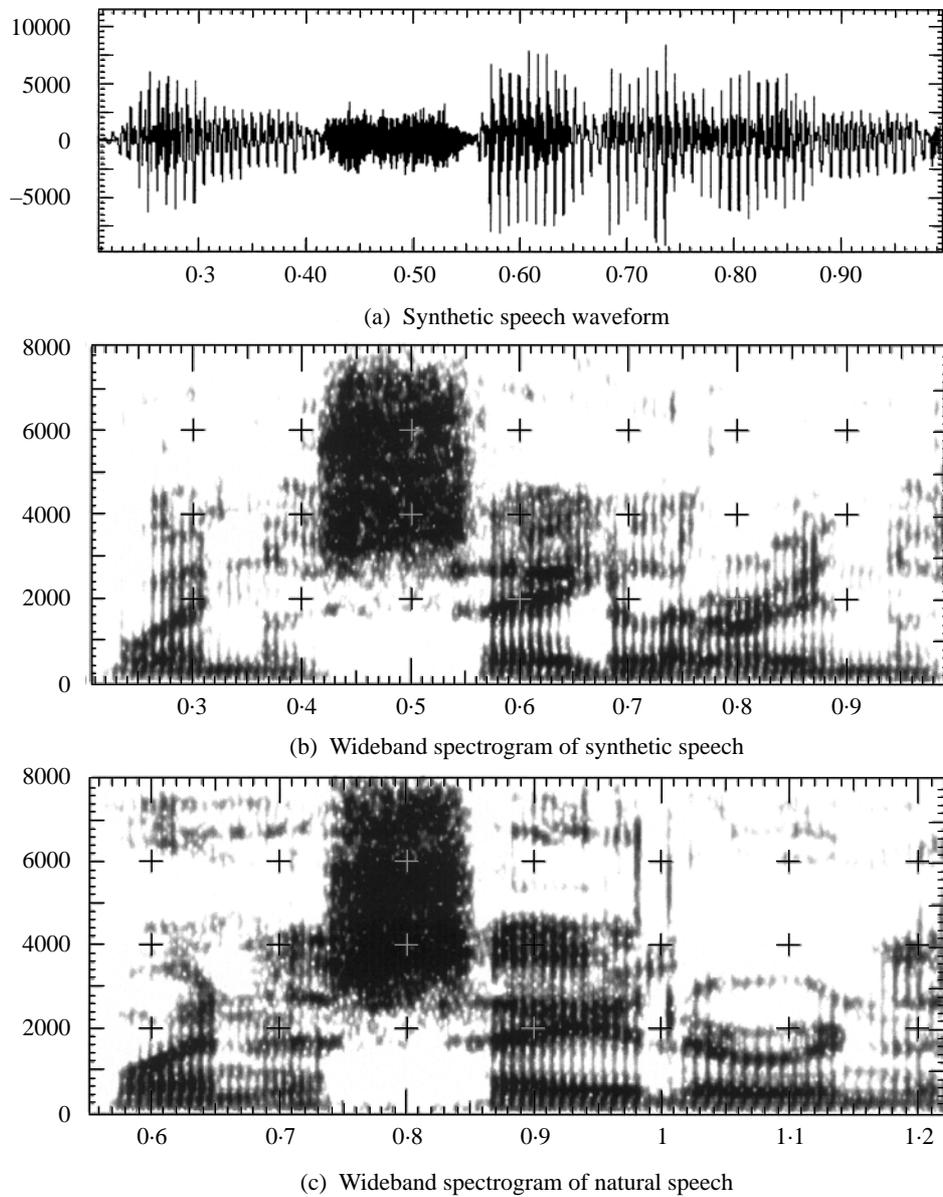
formant continuity is not as good. Although the introduction of syllabic clustering did improve the synthesis of many /l/s, the main defect in the speech corresponding to this region of the spectrogram seems to be a burbling sound through the /l/ of “sailor”. When looking at the formant structure in the spectrogram it is perhaps surprising that the synthetic speech sounds as good as it does. This phenomena has also been observed with other synthetic speech.

### 7.2. Modified rhyme tests

Modified Rhyme Tests (House, Williams, Hecker & Kryter, 1965) were conducted throughout the course of this research both as a measure by which to compare the system to others, and to obtain diagnostic information about the failings of the system. These tests usually used six listeners, who had in general performed the tests before, for reasons of speed and convenience. However, for the final system it was desirable to obtain an MRT score which could realistically be compared with those for other systems listed in Logan, Greene and Pisoni (1989) (see Table 7.2). For this system, trained on the M2 database, large scale MRTs were therefore conducted using 36 inexperienced listeners. The listeners used were gathered by advertising for native British English speakers, with no history of hearing problems, and no previous experience of working with synthetic speech. The experimental set-up was also very similar to that used by Logan *et al.* (1989). Six speech files were prepared, each of which consisted of a brief introduction to the task, followed by the test words separated by 4 second intervals of silence. The subjects were provided with an answer sheet, and asked to put a line through whichever of the six words on the answer sheet they thought they had heard for each test word played. They were asked to provide only one answer in each case, and to guess if they were not sure.

The MRT error rate obtained was 5.00%, with a standard error of 0.47%. The most frequently occurring word final errors involved the voiced plosives /d/ and /g/, and nasals, which together accounted for over 79% of word final errors. The word initial errors were more distributed, with errors in the identification of both voiced and unvoiced plosives, /l/ and /h/ all occurring frequently.

Many of the word final errors with plosives occurred between words which differed only in voicing. For example “bead” was often recognized as “beat”, and “pig” as “pick”, etc. This was because although the final plosive was often well produced, the length of the vowel was often inappropriate; an unvoiced plosive following a vowel should shorten that vowel



**Figure 1.** A synthetic speech waveform (a) and wideband spectrograms (b,c) of the sentence fragment “When a sailor in a . . .”. The synthetic speech (a,b) was produced using a system trained on the M2 database. The natural speech (c) was spoken by the speaker used in the M2 database.

(Klatt, 1987). Clustering questions did exist in the question list which could have split the vowel nodes by asking about the voicing of the following plosive, but they were often not used during tree building. The problem was perhaps that, with the acoustic analysis used, the acoustic difference between such vowels was small, despite the fact that their durations

differed considerably. A possible solution to this problem would be the use of an improved duration model, in which durations were clustered directly.

The use of MRTs to test the current system involves a mismatch between the training data, which was continuous speech, and the test data, which is isolated words. Isolated words are in general much slower than their continuous equivalents. The extra time available may mean that human speakers produce isolated words somewhat differently from how they produce the same words in continuous speech, and therefore that simply slowing down the continuous training speech to produce isolated words with the synthesizer may not be ideal. In addition, every isolated word contains a transition both to and from silence, whereas in the training data such transitions are relatively rare, since they only occur at phrase boundaries. These concerns suggest that the system may always have problems with synthesizing isolated words, unless it is specifically trained on an isolated word database, in which case it might do much better on an isolated word test like the MRT.

### 7.3. Discussion

As demonstrated by the MRT score described in the last section the system produces highly intelligible speech. In addition, as can be heard by listening to the audio examples available from Donovan (1996b) the speech sounds remarkably natural. The system's good performance is due to both the use of a high quality synthesis scheme and the automatic selection of subword units. An explanation as to why the HMM state-based approach selects such a good set of subword units is attempted below.

The clustering process used by the system is similar to those of other speech synthesis systems developed in recent years discussed in Section 2. One major difference here is that the clustering is state based instead of phone based. State-based clustering has been shown to outperform phone-based clustering for speech recognition (Young *et al.*, 1994; Odell, Woodland & Young, 1994), and similar gains may therefore be expected to occur in speech synthesis. The advantages arise because HMM states can be represented by a single feature vector, and thus lend themselves to clustering more directly than longer segments. Furthermore, the individual HMM states within a phone model can be clustered independently, which enables better use to be made of a given amount of training data. The smoothness of the formants of the synthetic speech is undoubtedly also due in part to the use of first- and second-order differential parameters in the feature vector representing each state. This means that states can be characterized by their dynamic features in addition to their static features.

The HMM-based segmentation process is similar to previous attempts discussed in Section 2. In this system, however, not only is the HMM system more sophisticated, but the synthesis units are segmented using the models created by the clustering process. Using the clustered state models to perform the segmentation enforces a large degree of consistency throughout the entire segmentation process. Boundaries between states will be well defined when those states are very different, when accurate segmentation matters most for synthesis, and less well defined when the states are more similar, which is when it matters least for synthesis. Furthermore, even if boundaries are located inaccurately, as judged by a human observer, provided the boundaries are consistently placed the "error" will be undone in synthesis. That is, segments which are likely to appear adjacent to each other during synthesis are likely to come from states which were often adjacent during construction. This means that the segments used in synthesis are likely to concatenate smoothly, because the boundaries of the states involved are likely to have been segmented consistently with each other. Thus to some extent, consistency is more important than accuracy. However, new state sequences can

occur during synthesis which were not present in the original database, and so accuracy is also important.

The speech synthesized by the system is not perfect. In addition to the duration problems mentioned in Section 7.2, the speech also suffers from occasional serious formant discontinuities and from occasional glitches caused by segmentation errors in the original database. Even when all these problems are absent, and the synthetic speech is generated using prosody transplanted from a human version of the same utterance, the speech still has a slightly uncertain, hesitant, character to it, which is thought to be the cumulative effect of all the tiny formant discontinuities between adjacent segments. A method which could potentially solve all of these problems is discussed in the next section.

## **8. Future work**

The segment selection algorithm used in the system, described in Section 5.1, works reasonably well, but is actually quite simplistic. A more sophisticated approach would be to make all the segments in each state available for use at synthesis, and to select between them during synthesis to produce the optimal segment sequence for the utterance to be synthesized. This could be achieved using a dynamic programming (d.p.) algorithm to minimize a suitably chosen cost function. Ideally, this cost function would specify the relative importance of spectral concatenation discontinuities between consecutive segments and the deviance of each selected segment from the required fundamental frequency, energy, duration, and state acoustic mean vector. It might take the form of a set of cost curves specifying, for each factor, and for a particular synthesis technique, what cost was associated with what degree of deviation. For example, when using TD-PSOLA synthesis, the duration cost curve would rise sharply as the segment duration dropped below half the required duration.

Research by Black and Campbell (1995) seems to indicate that human listeners prefer to hear speech which is smooth at the expense of acoustic accuracy, rather than vice versa. It is therefore likely that, if properly optimized, the cost function discussed above would strongly encourage concatenation smoothness during synthesis. Since no segments are likely to concatenate more smoothly than those which were originally adjacent, the algorithm should therefore encourage the use of a large number of adjacent segments in synthesis. Thus, the result would be a system which effectively concatenated variable length units, using longer units wherever they were available and it was advantageous to do so, whilst keeping the underlying state-based approach when longer units were not available. For each utterance, the d.p. algorithm would search over all variable-length segment sequences allowed by the decision trees, and find the sequence which was optimal in terms of the cost function, without an explicit list of variable-length segments being required.

The system described above would require the entire training database to be available during synthesis, which is impractical (at the time of writing) for many applications. Some form of pre-selection would therefore be required to establish which segments were most worth storing for a given inventory size. One possible method would be to use the cost curves to cluster the segments comprising each state into a number of self-similar subgroups, and then pre-select only one segment to represent each subgroup. Another would be to synthesize a large amount of test speech, and select the segments most frequently used.

The duration and energy prediction methods used could also be improved upon. Directly clustering duration and energy data instead of predicting them as a side-effect of the acoustic clustering would probably be advantageous. In addition cross-validation techniques (Breiman, Friedman, Olshen & Stone, 1984) should be used to determine the optimum depth of the

decision trees in terms of their ability to predict the durations or energies of held out data. Including other, more linguistically motivated, clustering features may also be advantageous. In addition, the decision tree approach fails to take advantage of the consistent effects of, and interactions between, different contextual factors when predicting durations for contexts not seen in training, and is therefore perhaps not the best model to use for duration prediction, and maybe energy prediction, at all. Considering alternative models, such as the Sums-of-Products model (van Santen, 1994), which attempt to model these consistencies, may therefore bring further improvements.

## 9. Conclusions

A speech synthesis system has been developed which uses a set of decision-tree state-clustered HMMs to select and segment a set of HMM-state-sized waveform segments to be its synthesis units. The system can synthesize natural sounding, highly intelligible speech, in a monotone, from a word string specification of known pronunciation. The segmental intelligibility of the speech has been measured using large scale Modified Rhyme Tests, and an error rate of only 5.0% obtained. The system can be retrained on a new voice in less than 48 hours, and has been successfully trained on four voices.

The results achieved by the system demonstrate the validity of this approach to speech synthesis, and the possibilities which exist for further development suggest that this approach holds considerable potential for the future.

Thanks to Tina Burrows and Patricia for recording databases, and to all the members of Cambridge University Engineering Department's Speech Vision and Robotics group who performed listening tests during system development. This work made extensive use of the hidden Markov model toolkit HTK developed at Cambridge University by Steve Young, Phil Woodland, and others. It also made use of the British English Example Pronunciations Dictionary developed at Cambridge University by Tony Robinson and others. The Engineering and Physical Sciences Research Council supported R.E.D. during this work.

## References

- Allen, J., Hunnicutt, M. S. & Klatt, D. (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge.
- Bahl, L. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D. & Picheny, M. A. (1991). Decision trees for phonological rules in continuous speech. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '91*, Toronto, pp. 185–188.
- Bigorne, D., Boëffard, O., Cherbonnel, B., Emerard, F., Larreur, D., Le Saint-Milon, J. L., Metayer, I., Sorin, C. & White, S. (1993). Multilingual PSOLA text-to-speech system. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '93*, Minneapolis, Volume 2, pp. 187–190.
- Black, A. W. & Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. *Proceedings on Eurospeech '95*, Madrid, pp. 581–584.
- Boëffard, O., Cherbonnel, B., Emerard, F. & White, S. (1993). Automatic segmentation and quality evaluation of speech unit inventories for concatenation-based, multilingual PSOLA text-to-speech systems. *Proceedings on Eurospeech '93*, Berlin, pp. 1449–1452.
- Boëffard, O., Miclet, S. & White, S. (1992). Automatic generation of optimized unit dictionaries for text-to-speech synthesis. *Proceedings on the International Conference on Speech and Language Processing '92*, Banff, pp. 1211–1214.
- Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove, CA.

- Brugnara, F., Falavigna, D. & Omologo, M. (1992). A HMM-based system for automatic segmentation and labelling of speech. *Proceedings on the International Conference on Speech and Language Processing '92*, Banff, pp. 803–806.
- Charpentier, F. J. & Stella, M. G. (1986). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '86*, Tokyo, pp. 2015–2018.
- Courbon, J. L. & Emerard, F. (1982). SPARTE: a text-to-speech machine using synthesis by diphones. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '82*, Paris, pp. 1597–1600.
- Dixon, N. R. & Maxey, H. D. (1968). Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio and Electroacoustics*, **AU-16**, 40–50.
- Donovan, R. E. (1996a). *Trainable Speech Synthesis*. PhD Thesis, Cambridge University Engineering Department Available from URL: [http://svr-www.eng.cam.ac.uk/People/Ex\\_Students/red/Personal.html](http://svr-www.eng.cam.ac.uk/People/Ex_Students/red/Personal.html), or by anonymous ftp to svr-ftp.eng.cam.ac.uk
- Donovan, R. E. (1996b). *PhD Thesis Audio Examples*. Cambridge University Engineering Department. Available from URL: [http://svr-www.eng.cam.ac.uk/People/Ex\\_Students/red/Personal.html](http://svr-www.eng.cam.ac.uk/People/Ex_Students/red/Personal.html), or by anonymous ftp to svr-ftp.eng.cam.ac.uk
- Donovan, R. E. & Woodland, P. C. (1995). Automatic speech synthesizer parameter estimation using HMMs. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '95*, Detroit, pp. 640–643.
- Fourcin, A. J. & Abberton, E. (1971). First applications of a new laryngograph. *Medical & Biological Illustration*, **21**, 172–182.
- Giustiniani, M. & Pierucci, P. (1991). Phonetic ergodic HMM for speech synthesis. *Proceedings on Eurospeech '91*, Geneva, pp. 349–352.
- Hallahan, W. I. (1996). DECTalk software: text-to-speech technology and implementation, URL: <http://www.europe.digital.com/info/DTJK01>.
- Hauptmann, A. G. (1993). SpeakeZ: a first experiment in concatenation synthesis from a large corpus. *Proceedings on Eurospeech '93*, Berlin, pp. 1701–1704.
- House, A. S., Williams, C. E., Hecker, M. H. L. & Kryter, K. D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158–166.
- Hunt, A. J. & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '96*, Atlanta, pp. 373–376.
- Itoh, K., Nakajima, S. & Hirokawa, T. (1994). A new waveform speech synthesis approach based on the COC speech spectrum. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '94*, Adelaide, Volume 1, pp. 577–580.
- Iwahashi, N., Kaiki, N. & Sagisaka, Y. (1992). Concatenative speech synthesis by minimum distortion criteria. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '92*, San Francisco, pp. II-65–II-68.
- Klatt, D. H. (1982). The Klattalk text-to-speech conversion system. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '82*, Paris, pp. 1589–1592.
- Klatt, D. H. (1987). Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, **82**, 737–793.
- Lee, K-F., Hayamizu, S., Hon, H-W., Huang, C., Swartz, J. & Weide, R. (1990). Allophone clustering for continuous speech recognition. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '90*, Albuquerque, pp. 749–752.
- Lernout, & Hauspie, (1999). URL: <http://www.lhs.com/speechtech/pcmmdevtools/tts.asp>.
- Ljolje, A. & Riley, M. D. (1991). Automatic segmentation and labeling of speech. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '91*, Toronto, pp. 473–476.
- Ljolje, A. & Riley, M. D. (1993). Automatic segmentation of speech for TTS. *Proceedings on Eurospeech '93*, Berlin, pp. 1445–1448.
- Logan, J. S., Greene, B. G. & Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, **86**, 566–581.
- Moulines, E. & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, **9**, 453–467.
- Nakajima, S. (1993). English speech synthesis based on multi-layered context oriented clustering. *Proceedings on Eurospeech '93*, Berlin, pp. 1709–1712.
- Nakajima, S. & Hamada, H. (1988). Automatic generation of synthesis units based on context oriented clustering. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '88*, New York, pp.

- 659–662.
- Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD Thesis, Cambridge University Engineering Department. Available by anonymous ftp to svr-ftp.eng.cam.ac.uk
- Odell, J. J., Woodland, P. C. & Young, S. J. (1994). Tree-based state clustering for large vocabulary speech recognition. *Proceedings on the International Symposium on Speech, Image Processing, and Neural Networks*, Hong Kong, pp. 690–693.
- Olive, J. P. (1990). A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds. *Proceedings on the ESCA Workshop on Speech Synthesis, AuTRANS*, Grenoble, pp. 25–29.
- van Santen, J. P. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, **8**, 95–128.
- Sharman, R. A. (1994). Concatenative speech synthesis using sub-phoneme segments. *Proceedings on the Institute of Acoustics*, **16**, 367–374.
- Sorin, C. (1994). Towards high-quality multilingual text-to-speech. *Proceedings on the CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology*, Munich, pp. 53–62.
- Taylor, P. A. & Isard, S. D. (1991). Automatic diphone segmentation. *Proceedings on Eurospeech '91, Genoa*, pp. 709–711.
- Tokuda, K., Kobayashi, T. & Imai, S. (1995a). Speech parameter generation from HMM using dynamic features. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '95, Detroit*, pp. 660–663.
- Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. & Imai, S. (1995b). An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. *Proceedings on Eurospeech '95, Madrid*, pp. 757–760.
- Wang, W. J., Campbell, W. N., Iwahashi, N. & Sagisaka, Y. (1993). Tree-based unit selection for english speech synthesis. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '93, Minneapolis*, Volume 2, pp. 191–194.
- Woodland, P. C., Odell, J. J., Valtchev, V. & Young, S. J. (1994). Large vocabulary continuous speech recognition using HTK. *Proceedings on the International Conference on Acoustic, Speech and Signal Processing '94, Adelaide*, Volume 2, pp. 125–128.
- Young, S. J., Odell, J. J. & Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. *Proceedings on the ARPA Workshop on Human Language Technology, Plainsboro, New Jersey*, pp. 307–312.

(Received 6 May 1998 and accepted for publication 9 April 1999)