A biphone constrained concatenation method for diphone synthesis

H. Timothy Bunnell, Steven R. Hoskins, and Debra M. Yarrington

Speech Research Laboratory, duPont Hospital for Children and University of Delaware
1600 Rockland Rd, Wilmington, DE 19803 USA

ABSTRACT

Diphone concatenation [1] has the advantages of simplicity and a relatively small database of speech when compared to other concatenative synthesis methods (e.g., [2]). However, diphone concatenation faces two notable problems. The first is coarticulation which extends beyond the scope of a single diphone and entails some degree of contextual mismatch for virtually any diphone in at least some concatenation contexts. The second problem, which stems from the first, is computational. It is the problem of selecting, from a specific speech corpus, an optimal instance of each diphone to achieve the least amount of temporal and spectral distortion in the broadest set of concatenation contexts (e.g., [3]).

We present a variant of diphone synthesis which addresses both problems by (a) allowing multiple tokens of diphones where needed to accommodate the effects of coarticulation, and (b) postponing diphone selection until synthesis when optimization can be constrained by known contextual factors. This method, termed Biphone Constrained Concatenation (BCC), has been implemented for use in the ModelTalker TtS system [4]. Comparisons of speech synthesized using BCC versus speech synthesized using pure diphone concatenation indicate clear improvements in naturalness for the BCC method. However, our listening experiments also demonstrated some increase in consonant confusions for the BCC method due to uncontrolled durational factors.

1. INTRODUCTION

Concatenative synthesis is capable of producing very natural sounding speech which, moreover, closely models the voice quality of the speaker who recorded the corpus from which the concatenation units are drawn. Of the various concatenation approaches possible (see [5] for a review), diphone concatenation [1] has the advantages of being simple to implement and requiring a relatively small database of speech. This makes diphone concatenation an attractive approach for applications which must share limited system resources with other processes and where it is undesirable or impossible to record large speech corpora to support the synthesis. For instance, synthesizers used in dialogue systems may need to share system resources with multiple processes, many of which are inherently resource-demanding. People with degenerative neurological disorders are excited about the possibility of "capturing" their voice for use in an augmentative communication device when their disease progresses to the point that their own voice is lost. However, these individuals may not be able to record large speech databases. Diphone concatenation may be more practical for these individuals.

Despite its advantages, diphone concatenation faces two notable problems. The first problem stems from the presence of coarticulation, which entails that the influence of segments external to a diphone may be present in the acoustic structure of the diphone; diphones cannot be viewed as context independent units. Thus, it is generally not possible to select a diphone from one context and insert it into another context without introducing some degree of distortion. In the worst case, the distortion amounts to abrupt spectral discontinuities at diphone boundaries. However, even if there are no obvious spectral discontinuities, distortion may still be present due to coarticulatory effects. Several classic reaction time studies have shown that perception is disrupted by the presence of misleading coarticulatory effects in speech, even when there is no audible discontinuity [6][7][8]. These studies can be interpreted to suggest that listener's expectations regarding coarticulation can be violated by a mismatch between the coarticulatory effects present in a diphone and the effects predicted from its context. Typically, coarticulatory effects have been addressed by adding polyphones to the unit inventory to cover heavily coarticulated phonetic sequences. Of course, this has the disadvantage of increasing the inventory size.

The second problem is computational in nature and stems from a desire to minimize the effects of coarticulation. Typically, diphones are selected from a corpus of speech containing multiple instances of many of the diphones. From these multiple instances, a single instance of each diphone is selected for the synthesis inventory, and the computational problem is that of selecting the precise diphone instance which will minimize distortion effects in the greatest number of possible concatenation contexts. Earlier work in our laboratory approached this problem by finding the set-wise minimum for distortion among potential diphone boundaries within phonemes [9]. The process located the best phoneme candidates (and boundaries within them) and then selected diphones to minimize the combined phoneme distortion measures. This process was computationally expensive and did not, in fact, guarantee a globally optimal solution since it minimized distortion at the phoneme rather than diphone level; it is possible that the "best" diphone is not the one derived from the best scoring pair of phonemes.

To summarize, computational complexity makes it difficult to select an optimal set of diphones, while the presence of coarticulation means that even if the computation is successful, it will be at least partly futile. To address both these problems, we present a variant of diphone synthesis which (a) postpones diphone selection until synthesis when optimization can be constrained by known contextual factors, and (b) allows multiple tokens of diphones to accommodate the effects of coarticulation. This method is termed Biphone Constrained Concatenation

(BCC). Although the concatenation unit in BCC is the diphone, the primary unit of speech storage and retrieval is the biphone (i.e., two complete adjacent phonetic segments). The BCC method shares, with non-uniform unit selection methods, the use of a dynamic programming search to locate the best set of concatenation units for synthesis of a specific utterance. It differs from non-uniform unit selection methods in the size and structure of the speech database from which it draws units for concatenation, and in that the units are specifically diphones.

The BCC method has been implemented for use in the ModelTalker TtS system [4]. In the following, we describe the BCC method in detail and describe initial experiments which compare both the intelligibility and the naturalness of speech synthesized with the BCC method to synthetic speech from the previous diphone-based phoneme to sound component in ModelTalker.

2. BIPHONE CONSTRAINED CONCATENATION

BCC generates speech for a given sequence of phonemes by concatenating diphones drawn from an indexed speech database. Unlike most diphone synthesis systems which store a single instance of each diphone in the inventory, BCC stores multiple instances of biphones (phoneme pairs) from which diphones can be extracted during synthesis. The selection of specific diphone instances at synthesis time minimizes temporal and spectral distortion for the phoneme sequence to be realized acoustically. This approach represents a compromise between the small size of regular diphone synthesis systems (e.g., [10]). and the fluent quality of large-scale unit concatenation systems [5]. It further allows the developer of the speech database to select a tradeoff between size and quality of synthetic speech, and even allows specific utterances to be stored and reproduced verbatim.

The two factors most important to understanding how BCC functions are the construction of the speech database, and the search strategy used during synthesis. Each of these is described below.

2.1. Database construction

Database construction begins with the recording and labeling of an appropriate speech corpus. We discuss first the corpus structure and then the recording and labeling procedures.

2.1.1. Corpus Structure

As part of a systematic study of the best corpus design for diphone synthesis, several lists of utterances have been developed. These lists are algorithmically generated to trade compactness for control of prosodic and segmental context. The most compact list of utterances is constructed of short syntactically acceptable but semantically anomalous phrases, each phrase having a maximum of seven syllables (e.g., *fatuous bows; stilted English stunts)*. A less compact list which is better controlled for prosodic structure is also made of syntactically acceptable English word sequences with a maximum of three syllables (e.g., *rose shop; fizz pot*). The least compact and most carefully controlled list contains only a few real words which derive fortuitously from nonsense word forms intended to minimize both prosodic and segmental variation (e.g., *ah sus ah*). For "complete" coverage (i.e., > 99% of the needed diphones), about 500 7-syllable, 750 3-syllable, and 1200 nonsense phrases must be recorded.

In addition to the complete inventory lists, we have identified sublists which (a) provide 90% coverage of all needed diphones, and (b) complete coverage of a set of semantically anomalous sentences often used in this laboratory for intelligibility testing. It was from these latter small inventories that databases were built for synthesis experiments described here.

The primary intended use of the ModelTalker system is in augmentative communication devices, and one potential population of users are those with neurodegenerative disease who wish to record their own voice for use in a synthesizer. Among this population, it may be necessary to look for trade-offs between corpus size and synthetic speech quality. These lists allow us to explore this domain.

2.1.2. Recording and Labeling

All recordings are made with the talker seated in a sound-attenuated room, facing the monitor of a PC running Windows 95/NT. The talker wears a head-mounted microphone and headphone set (Sennheiser HMD-410), and electroglottograph (EGG) transducer (Glottal Enterprises model MC2-1). Digitization is accomplished with a Turtle Beach Fiji sound card for Windows. Speech is normally digitized at a 16000 kHz sampling rate.

The software for recording and labeling is called InvTool. InvTool accepts a list of utterances to be recorded along with phonetic transcriptions for the utterances and produces a set of waveform files (one for each utterance) with corresponding acoustically aligned phonetic labels. For the waveform files, InvTool uses an extension of the standard RIFF waveform file format which allows the label information to be stored in the waveform file along with the speech data.

For each utterance to be recorded, InvTool presents an auditory prompt—a prerecorded example of the utterance—along with a text display of the English gloss for the utterance. After being prompted, the talker initiates recording with a mouse click, speaks the utterance, and then terminates recording with another mouse click. When the talker terminates recording, an immediate waveform and spectrogram display of the recording appears, and from this display, the utterance (sans initial and trailing silence) is selected for processing. The processing entails alignment of phonetic labels using the HMM alignment tool described in [11] and pitch tracking based on information in the EGG and/or acoustic speech signal.

Although the InvTool phonetic labeler is at least as accurate as commercially available systems, the labels assigned by InvTool are not error free. The temporal resolution of the labeler in InvTool is 10 msec, and generally, between 10 and 15 percent of the labels assigned by InvTool fall outside the 10 msec range of acceptable positioning and need to be hand corrected. Moreover,

some acoustic features—notably stop bursts—require better than 10 msec resolution. For these, additional software has been developed which places segment boundaries involving stop and affricate bursts with necessary precision based on the initial boundary assignment of the InvTool labeler.

Once speech has been labeled, pitch tracked, and corrected, a biphone database is constructed consisting of a tagged index file and a synthesis data file. The tagged index file (hereafter IDX file) contains an entry for each biphone in the database and describes the biphone on a pitch period by pitch period basis. Specifically, each pitch period (or short voiceless epoch) is described by its duration, voicing status, spectral features (as a codeword described below), and a pointer to its stored representation in the synthesis data file. The synthesis data file (hereafter DAT file) itself contains the parameters needed to synthesize each pitch period. For the present version of the synthesizer, this database is windowed PSOLA data.

To obtain codewords to describe the acoustic structure of each pitch period or voiceless epoch, the speech data are analyzed pitch-synchronously. A Bark cepstrum analysis is used to derive acoustic feature vectors for each epoch in the database and the feature vectors are subjected to vector quantization to produce the codebook which is used to classify each epoch and to characterize the spectral distance between each pair of epochs. Our experimental inventories use 256-vector codebooks for the classification. A vector distance matrix is computed to provide a lookup table for a Bark Cepstral distortion measure between every codebook vector. The distortion measure used for this is:

$$D_{ij} = \sqrt{\sum \left(C_{ik} - C_{jk}\right)^2} \qquad (1)$$

Where D_{ij} is the distortion between the i^{th} and j^{th} codebook vectors. This 256 by 256 element distance matrix is used to estimate inter-frame (pitch period) differences during synthesis. Using this lookup table for distances avoids having to store cepstral feature vectors and perform the more costly (though more accurate) distortion computations at the time of synthesis.

2.2. Synthesis

The ModelTalker system synthesizes speech in intonation phrase sized chunks. In the synthesis process, each new phoneme within the target phrase is comprised of a section from a predecessor diphone and a successor diphone. These diphones are extracted from predecessor and successor biphones by searching among all possible candidate biphones for the diphone region that will minimize a cost function calculated over the phrase to be synthesized. To do this, a dynamic programming strategy is used which accumulates a cost function for each possible diphone boundary over the phrase. At that point, the sequence of boundaries which defines the least costly series of diphones is determined. This becomes the diphone series to be used for the phrase. This is further illustrated in Figure 1, which shows how costs are computed for the boundary within a single phoneme. This phoneme is to be constructed from data in the trailing half of one of three predecessor diphones and the leading portion of one of three successor diphones. (Note that the number of predecessor and successor diphones need not be

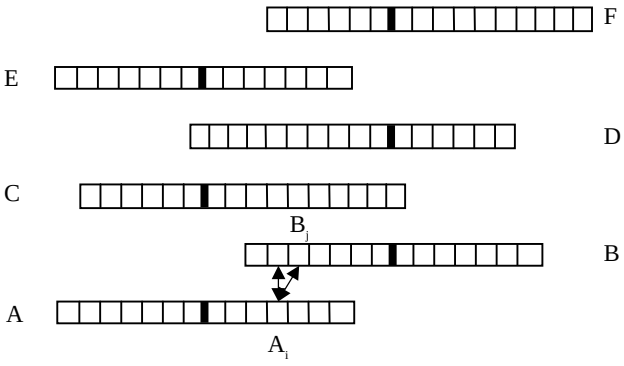

Figure 1. Boundary search among candidate successor biphones {B,D,F} and predecessor biphones {A,C,E}. The splice cost is added to the path cost for the predecessor, and retained as the path cost for the successor diphone. Splice costs are determined by spectral and temporal distortion measures as described in the text.

equal, and that candidate phonemes are computed for all combinations of predecessors and successors). The particular biphones to be used, and the exact boundary locations within them are to be determined in part by a local cost function which comprises three terms, a spectral distortion measure, and two temporal distortion measures. The spectral distortion measure is:

$$S_{ij} = D_{A_i B_j} + D_{A_i B_{j-1}} \qquad (2)$$

where the A_i and B_j denote the vectors associated with frame *i* in predecessor biphone *A* and frame *j* in successor biphone *B* respectively, and *D* is the distance measure as previously described.

The temporal distortion measures are defined for (a) specific pairs of predecessor and successor biphones, and (b) specific diphone boundary locations within a pair of biphones. The first of these is constant for all boundary locations within a pair of potentially conjoining diphones and is defined as:

$$T1_{AB} = \left(\left|A_{dur} - B_{dur}\right|\right) / \left(A_{dur} + B_{dur}\right) \qquad (3)$$

where A_{dur} and B_{dur} are the durations of the overlapping phonemes in predecessor biphone *A* and successor biphone *B* respectively. This measures the duration contrast between the overlapping phoneme durations of the potential predecessor and successor biphones and discourages selection of diphones from segments having greatly different durations.

The second temporal distortion measure is boundary location specific and is minimal when the duration of the segment that would be created by selecting a specific boundary for conjoining two diphones would result in a segment that is the same duration as the segment in the successor biphone being considered. That is:

$$T2_{ij} = \left|1.0 - \left(A_{dur}^{0-i} + B_{dur}^{j-t}\right) / B_{dur}\right| \qquad (4)$$

where *0-i* denotes the portion of the segment from its beginning to the i^{th} epoch, and *j-t* denotes the portion of the segment from the j^{th} epoch to its last epoch at frame *t*. The sum of these two partial segments is the duration of the resulting synthetic segment and the deviation of this duration from that of the

segment in the successor biphone is its cost. The rationale behind this temporal distortion measure is that English vowel durations are more strongly predictive of the voicing of subsequent than prior consonants, and consonant durations are less predictive of surrounding vowel identity in either direction. Thus, to the extent that a bias toward the duration of either segment should be exerted, it should logically go toward preserving the duration of the segment with which future context is coarticulated.

To illustrate the differences observed in synthesis from one corpus, Figure 2 shows excerpts of the same semantically anomalous sentence synthesized with both the BCC method and the standard ModelTalker automatically extracted diphone concatenation. The full sentence, available on the CD-ROM as [SOUND R63_01.WAV] for the original ModelTalker sentence (top portion of Figure 2), and [SOUND R63_02.WAV] as the BCC version, was *The bait is reaping the coo* (bottom portion of Figure 2). Note in this figure the "scalloped" formant structure in the first word top panel and the discontinuous F2 of the last word in the top panel. Neither problem appears in the BCC version of the same sentence as shown in the bottom panel.

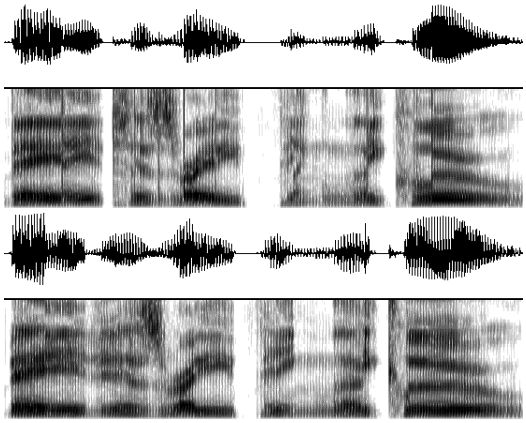

Figure 2. Waveform and spectrogram displays for the sentence *"…bait is reaping the coo."* Top version synthesized from automatically extracted diphones. Bottom version synthesized using the BCC method.

While these differences appear compelling, to determine how well the BCC method works compared to the standard ModelTalker diphone concatenation, two evaluation experiments have been carried out. These experiments compare the intelligibility and naturalness of synthetic speech synthesized with the BCC method to that of speech synthesized using the previous ModelTalker diphone concatenation. Both of these experiments used a set of semantically anomalous sentences that we have used in the past for evaluating the intelligibility of both synthetic speech and disordered natural speech (e.g. [11][12]). In the first experiment, intelligibility was assessed by totaling the number of consonant and vowel confusions observed among groups of closed response target items. This experiment also recorded fluency ratings from listeners. The second experiment used the same sentence set in a forced choice paired comparison task where listeners selected which member of a sentence pair was more natural sounding. Each of these experiments is described below in greater detail.

3. EXPERIMENT I - INTELLIGIBILITY

3.1. Methods

3.1.1. Stimuli

The same 74 sentences were synthesized with each of two versions of the ModelTalker synthesizer. The versions differed in the phoneme to sound module. One version used the original ModelTalker diphone concatenation module (the DIPHONE version), the other used the new BCC phoneme to sound module (the BCC version). The waveform processing to control f0 in these two versions was very similar, but not identical. The DIPHONE module used a pitch synchronous method in which overlap blending of pitch periods involved asymmetrical windows; the BCC version used standard PSOLA overlap blending with symmetrical windowing.

The speech databases, both DIPHONE and BCC, were derived from the same speech corpus recorded by a male talker. The corpus contained utterances of up to seven syllables; the selected utterances covered all the diphones needed to synthesize the set of test sentences.

The 74 sentences were of the form *The NOUN1 is VERBing the NOUN2* where NOUN1 and NOUN2 were drawn without replacement from a set of 74 monosyllabic nouns and the VERB was drawn without replacement from a set of 37 *ing*-form verbs. Each noun and verb in these sets is associated with a group of from four to six other nouns and verbs (also in the set) which differ minimally from it (e.g., din, tin, fin, pin, bin; beet, bit bet bat).

3.1.2. Listeners

The listeners were five employees of the Speech Research Laboratory. All were familiar with the experimental task and synthetic speech, but none had listened to the experimental sentences prior to the listening experiment.

3.1.3. Procedure

Listeners were seated in a sound dampened room facing a computer monitor which was used to display the sentence frames along with all response alternatives for each of the target words. The display further contained a set of fluency rating descriptors which allowed listeners to rate each sentence as MOST, MORE, OK, LESS, or LEAST fluent. A mouse was used to select responses.

On each trial, listeners heard a single sentence and were then required to select the target word from the alternatives for each of the three target positions as well as provide a fluency rating for the sentence. The trial ended and the next trial was initiated by entry of the fluency rating.

3.2. Results

Overall, listeners identified 93.9% of the target words correctly. For BCC concatenation, 93.4% of the words were correctly identified, and for DIPHONE concatenation, 94.4% of the words were correctly identified. This percentage difference amounts to an average of one more word identification error per listener. An examination of the specific errors disclosed that, of the 73 total errors for BCC concatenation, 28 involved instances wherein the stops /t/ and the glottal stop were heard as /d/ and /b/ respectively, and four instances in which the initial /n/ of *knowing* was identified as /g/ (i.e., *going*); the remaining errors were mostly voicing confusions. Of the total 62 errors in the DIPHONE set, errors were more broadly distributed. There were 9 instances in which /t/ was heard as a voiced consonant or glottal stop, 9 instances in which voiced stops were heard as voiceless stops, and the remaining errors primarily involved confusions among voiced consonants.

Despite the slightly greater intelligibility scores for the DIPHONE synthesized sentences, the fluency ratings favored BCC concatenation. With most fluent assigned a score of 1 and least fluent assigned a score of 5, the average fluency rating for BCC synthesized speech was 2.47 and for DIPHONE synthesized speech it was 3.39.

To explore this fluency difference further, the second experiment required listeners to directly compare sentences synthesized using the two methods by choosing which of two sentences presented as a pair sounded “more natural.”

4. EXPERIMENT II – NATURALNESS

4.1. Methods

4.1.1. Stimuli

All sentences used in Experiment I were also used in this experiment.

4.1.2. Listeners

The same five listeners participated in Experiment II.

4.1.3. Procedure

Sentence were presented in pairs with one member of the pair being a sentence synthesized using the BCC method and the other member of the pair being the same sentence synthesized with the DIPHONE method. The interstimulus interval was 500 msec for the sentences presented in a pair and the order of presentation was counterbalanced with BCC sentences being the first sentence in half the pairs and DIPHONE sentences first on the other half of the pairs.

On each forced-choice trial, the listeners were required to identify which of the two sentences sounded more natural.

4.2. Results

Overwhelmingly, (92.0% of the trials) listeners identified the BCC sentence as more natural sounding than the DIPHONE concatenation sentence.

5. DISCUSSION

The BCC method appears to produce substantially more natural sounding synthetic speech than did the previous diphone concatenation method used in the ModelTalker system. While the intelligibility of the BCC sentences was not greater in this experiment, there are several reasons to suspect that this result is not an inherent feature of BCC concatenation. The most significant of these has to do with the way our speech corpus was constructed for the present experiment. The corpus used in this experiment was originally recorded and labeled for use with our automatic diphone extraction software. Consistent with procedures used for automatic diphone extraction, the stop consonants of the present corpus were labeled as a single acoustic phonetic region, beginning at the closure of a preceding vowel and ending at the onset of a subsequent vowel. The crucial step of locating the release burst of the stop consonant to partition the stop into separate closure and release intervals was performed by the automatic diphone extraction software. However, the BCC software has been developed for acoustic phonetic labels which include separate closure and release labels as part of the phonetic inventory. To use the old corpus with the new BCC software, this labeling requirement was relaxed, but no additional logic was built into the BCC system to improve its functioning in the absence of separate closure and release labels. Consequently, it is likely that the very large proportion of voiceless-to-voiced stop errors observed for BCC synthesized speech resulted from inappropriately controlling the duration of the stop closure/release intervals.

A second factor may also have played a role in limiting the intelligibility of the BCC synthesized speech for the present study, specifically, the number and variety of diphones actually available for selection at synthesis time. The corpus used in the present study was carefully selected to cover the needed set of diphones with minimal redundancy. As a result, the majority of the diphones used in the present experiment could come from only one or two contexts in the corpus. In the case that there was only one context, the BCC method is constrained to choose exactly the same phonetic segments as the automatic diphone extractor from which to select a diphone (although the BCC-selected diphone boundaries may differ from those of the automatically extracted diphone). In many cases where multiple diphone instances were present in the corpus, not all instances were likely to be acoustically desirable. For instance, the sequence /sw/ used as a word-initial consonant cluster in several target words could be extracted (appropriately) from the word *sweep* or (inappropriately for this experiment) from the inter-word context *glass wife*. In these circumstances, the virtue of BCC concatenation to find the contextually best diphone was not exercised. At best, one can hope that the BCC method chose the same diphone as that selected by the automatic diphone extractor in this case.

Overall, the performance of the BCC concatenation method appears to hold great promise for maintaining some of the size/complexity constraints of pure diphone concatenation while allowing us to take advantage of features of non-uniform unit concatenation when it is desirable to do so.

One of the particularly interesting possibilities for this approach is the ability to store complete utterances as part of the BCC database. In augmentative communication, there has been continued concern about the impact of less natural synthetic speech, particularly in devices used by young children. As a result, some AAC devices have continued to provide the ability to store recorded words and phrases as separate from, and in some instances preferable to synthetic speech. Since it is possible to store entire words and phrases in a BCC concatenation database, users of AAC devices may enjoy the benefits of unconstrained synthesis and maintain a recorded quality speech capability (in the special case of utterances that have been stored fully within the database). This is because the search strategy would (presumably) return minimal error for a sequence involving an utterance within the database that was stored as spoken.

ACKNOWLEDGEMENTS

The authors wish to thank James Polikoff for his assistance in all aspects of this work. We acknowledge the support of NIDRR under grant number #H133E30010 and the Nemours Research Programs for continued programmatic support, and the Microsoft Corporation.

REFERENCES

[1] Peterson, G., Wang, W., and Siversten, E. (1958). Segmentation techniques in speech synthesis. . Journal of the Acoustical Society of America, 30, 739-742.

[2] Takeda, K., Abe, K., and Sagisaka, Y. (1992). On the basic scheme and algorithms in non-uniform unit speech synthesis. In G. Bailly, C. Benoit and T.R. Sawallis (eds.), Talking Machines: Theories, Models, and Designs. Amsterdam: Elsevier, 93-105.

[3] Conkie, A. D. and Isard, S. (1997). Optimal coupling of diphones. In Progress in Speech Synthesis, van Santen, J.P.H., Sproat, R.W., Olive, J.P., and Hirschberg, J. (eds.). Springer, New York, pp. 293-304.

[4] Bunnell, H.T., Hoskins, S. R., and Yarrington, D. M. (1998). The ModelTalker project: Software for diphone speech synthesis and automatic diphone extraction. University of Delaware, Computer and Information Sciences Technical Report, 98-13.

[5] Sagisaka, Y. (1995) Spoken output technologies. In Cole, R., Mariani, S., Uszkoreit, H., Zaenen, A. Zue, V. Survey of the state of the art in human languages technology. Center for the Spoken Language Understanding, Oregon Graduate Institute, Beaverton, Oregon. pp. 189-226

[6] Fowler, C.A. (1981b). Production and perception of coarticulation among stressed and unstressed vowels. Journal of Speech and Hearing Research, 46, 127-149.

[7] Martin, J.G. and Bunnell, H.T. (1981). Perception of anticipatory coarticulation effects. Journal of the Acoustical Society of America, 69, 559-567.

[8] Whalen, D. (1982). Perceptual effects of phonetic mismatches. PhD thesis, Yale University, New Haven, CT.

[9] Yarrington, D., Bunnell, H.T., and Ball, G. (1995). Robust Automatic Extraction of Diphones with Variable Boundaries. ESCA Eurospeech '95. 4th European Conference on Speech Communication and Technology. Madrid, Spain, September 1995, 1845-8.

[10] Moulines, E., and Charpentier, F. (1990) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9 (5/6), December, 453-67.

[11] Menendez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., and Bunnell, H.T. (1996) The Nemours Database of Dysarthric Speech. Proceedings of ICSLP 96, Philadelphia, USA, October 1996, 1962-5.

[12] Bunnell, H.T., and Polikoff, J.B. (in preparation). "Toward enhancement of dysarthric speech: Perceptual effects of temporal modifications."